This electronic thesis or dissertation has been downloaded from the King's Research Portal at https://kclpure.kcl.ac.uk/portal/

KING'S College LONDON

Algorithms and architectures for high resolution Sigma-Delta converters.

Magrath, Anthony J

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without proper acknowledgement.

END USER LICENCE AGREEMENT

Unless another licence is stated on the immediately following page this work is licensed

under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International

licence. https://creativecommons.org/licenses/by-nc-nd/4.0/

You are free to copy, distribute and transmit the work

Under the following conditions:

- Attribution: You must attribute the work in the manner specified by the author (but not in any way that suggests that they endorse you or your use of the work).
- Non Commercial: You may not use this work for commercial purposes.
- No Derivative Works - You may not alter, transform, or build upon this work.

Any of these conditions can be waived if you receive permission from the author. Your fair dealings and other rights are in no way affected by the above.

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

Algorithms and Architectures for High Resolution Sigma-Delta Converters

A Thesis Submitted to
The University of London
for the Degree of
Doctor of Philosophy

Anthony J. Magrath
King's College London
August 1996

BIBL. LOUVAIN

Abstract

Sigma-delta modulators are widely used for low-level digital-to-analogue (D-A) and analogue-to-digital (A-D) conversion in applications such as digital audio, communications, control and signal processing systems. The modulator utilises a technique known as noise shaping to achieve high baseband resolution with only a single-bit quantizer. The modulator may also be used in power D-A conversion, to achieve high power, high efficiency conversion, through the use of a two-level power switch.

This dissertation is concerned with enhancing the performance of sigma-delta modulators using a new class of algorithms, which operate by selective inversion of the state of the quantizer. This technique is termed bit-flipping. After a study of the modelling of sigma-delta modulators, three distinct applications of bit-flipping are considered.

Bit-flipping is used to increase the linearity of the modulator, which is otherwise limited by baseband tones that arise from periodic patterns in the quantizer output. It is shown that bit-flipping can increase the conversion linearity without the use of conventional random dither. Under certain conditions, lower noise penalties can be achieved than random dither. Furthermore the technique is well suited to A-D converter implementations.

It is next shown how bit-flipping may be used to adaptively reshape the noise spectrum introduced by the quantization process. This technique is used to reduce the baseband noise power at low signal levels, without seriously compromising stability at high signal levels. It is shown that improvements in dynamic range can be achieved, in comparison to standard modulators.

Finally, bit-flipping is used to improve the performance of sigma-delta based power D-A converters. The bit-patterns are modified to regulate and reduce the average pulse repetition frequency of the bitstream so that efficient power switching is possible. In comparison to the more common pulse-width modulation converters, the scheme offers high linearity and lower clock speeds.

To my parents

Acknowledgments

I would like to express my gratitude to my supervisor, Professor Mark Sandler for his guidance, encouragement and support throughout the project.

I am also grateful to several colleagues and members of staff for their help. In particular I would like to thank Simon Kershaw for his enthusiasm, for stimulating technical discussions and feedback. I am especially grateful to Ian Clark for undertaking the hardware implementation related to this work. For sharing their knowledge and insight I would like to thank Dr. Chris Dunn and Dr. Rod Hiorns. I would also like to thank members of the Circuits and Systems Research Group, including Julian Bean, Victor Bocharov, Gabriella Castellano, Panos Kudimakis, Georgi Petkov, Marc Price and Mike Waters for creating a friendly and lively working environment. I am indebted to Peter King, Talat Malik and Mustaq Mohammed for their infinite patience and technical support.

In addition, I am deeply grateful to my parents for their continual support and encouragement, and to my friends for keeping me sane.

Finally I would like to acknowledge King's College London, EPSRC and the Royal Academy of Engineering for providing financial support. I am also indebted to IMEC, Leuven, Belgium where some of the research was carried out.

Contents

List of Figures

List of Tables

List of Symbols

k	Discrete-time index
$b(k)$, $y_a(k)$	Input to Bit-flipping operator
$e(k)$, $e_l(k)$	Quantizer error
$e_w(k)$	Weighting filter output. Defined in figure 5.3
$f_e(k)$	Estimate of pulse repetition frequency at sample k
$n(k)$	AC component of quantizer input
$r(k)$	Dither source added to quantizer input
$u(k)$	Quantizer input
$v(k)$	Input to quantizer Q_B in dual quantizer model (figure 5.6)
$x(k)$	Modulator input
$y(k)$	Modulator output
$y_a(k)$, $y_b(k)$	Quantizer outputs in dual quantizer model (figure 5.6)
$y_{la}(k)$	One level look-ahead modulator output
$y_{2la}(k)$	Two level look-ahead modulator output
$G(z)$	Higher order weighting filter. Defined in figure 5.15
$H(z)$	Loop filter transfer function
$NTF(z)$	Noise Transfer Function. Defined in equation 2.3
$NTF_a(z)$, $NTF_b(z)$	Noise transfer functions in dual quantizer model. Defined in equations 5.21 and 5.23
$NTF_d(z)$	Noise transfer function of modulator with additional loop delays. Defined in equation 6.40
$NTF_K(z)$	Noise Transfer Function modified by quasi-linear gain. Defined in equation 2.4
$NTF_w(z)$	Noise transfer function associated with weighting filter. Defined in equation 5.30
$STF(z)$	Signal Transfer Function. Defined in equation 2.2

A_s Peak sinusoidal input amplitude

A_{max} Maximum stable amplitude

B Quantizer input bound. Defined in equation 4.3.
Also peak dither amplitude.

f_i Sinusoidal input frequency

f_l Frequency of dominant low frequency idle tone

f_h Frequency of dominant high frequency idle tone

f_p Average Pulse Repetition Frequency of $y(k)$. Defined in equation 6.3

$\hat{f_{pd}}$ Maximum pulse repetition frequency for DC input.
Defined in equation 6.13

$\hat{f_{ps}}$ Maximum pulse repetition frequency for sinusoidal input.
Defined in equation 6.16

f_s Nyquist sampling frequency

f_t Target pulse repetition frequency of bit-flipping algorithm

F_k Pulse repetition frequency constant. Defined in equation 6.22

K_n Quasi-linear quantizer noise gain

K_s Quasi-linear quantizer signal gain

m_s DC component of quantizer input

m_x DC input amplitude

m_y Mean output amplitude

L Oversampling Ratio

M Number of simulation samples

MQ Maximum quantizer input magnitude. Defined in definition D.1

MAQ Maximum active quantizer input magnitude. Defined in definition D.2

N Order of Noise Transfer Function

N_{min} Refer to definition 6.2

P Period of periodic bit sequence

P_b Baseband Noise Power (expressed in dB)

P_m Maximum NTF power gain for stability

P_n, P_f Noise Transfer Function Power Gain (expressed in dB)

$P_n(K_n)$ K_n-modified Noise Transfer Function Power Gain

P_1 Number of positive output samples in period P

P_2 Number of negative output samples in period P

S_a	Alternation constraint. Defined in section 6.3.2
t_r	Rise time of power switch
t_f	Fall time of power switch
T_r	Average transition rate of $y(k)$. Defined in equation 6.1
$\hat{T}_r$	Maximum transition rate of $y(k)$
δ	Peak dither amplitude
$\pm\Delta$	Quantizer output levels
ϵ	Upper bound on quantization error
$\mathcal{Q}\{.\}$	Quantizer transfer function
$sgn\{.\}$	Sign operator
σ_e^2	Variance of $e(k)$
σ_n^2	Variance of $n(k)$
σ_q^2	Noise Variance at modulator output
σ_w^2	Variance of $e_w(k)$. Defined in equation 5.1
θ	Normalised radian frequency
$E\{.\}$	Expectation operator
$V\{.\}$	Variance Operator
$p\{.\}$	Probability
$\|\|.\|\|_1$	One-norm i.e. sum of magnitudes
$\|\|.\|\|_\infty$	Infinity-norm i.e. maximum magnitude

List of Abbreviations

A-D	Analogue-to-Digital
ASIC	Application Specific Integrated Circuit
ASR	Active Selection Rate
ATR	Average transition rate
BF	Bit-Flipping
BFO	Bit-Flipping Operator
D-A	Digital-to-Analogue
DBF	Deterministic Bit-Flipping
DR	Dynamic Range
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FPGA	Field Programmable Gate Array
HF	High Frequency
IIR	Infinite Impuse Response
LA	Look-ahead
LCD	Limit Cycle Detector
LF	Low Frequency
MSA	Maximum Stable Amplitude
NTF	Noise Transfer Function
PDF	Probability Density Function
PRF	Pulse Repetition Frequency
PWM	Pulse-Width modulation
SNR	Signal-to-Noise Ratio
STF	Signal Transfer Function
VLSI	Very Large Scale Integration
WBF	Weighted Bit Flipping
$\Sigma\Delta$	Sigma-Delta

Chapter 1

Introduction

Since the invention of Pulse Code Modulation (PCM) by A.H. Reeves in 1937 [Ree65] the processing of signals in the digital domain has become widespread. Many advantages are gained from the use of digital signal processing (DSP), including noise immunity, robustness, complex functionality and flexibility. Over the last 40 years there has been an increasing shift towards the use of DSP to implement functions normally implemented in the analogue domain, fuelled by the continual reduction in cost and size of digital implementations. As a consequence, the analogue-to-digital (A-D) and digital-to-analogue (D-A) interface has become crucial to the operation of many electronic systems. Since it is possible to implement digital signal processors with very high precision, the system performance is ultimately limited by the precision of the interface. There has consequently been extensive research in the area of high resolution A-D and D-A converters, for applications including telecommunications, audio and video coding, control and instrumentation.

In this dissertation, particular emphasis is placed on the application of A-D and D-A converters to digital audio systems. The aim of digital audio is to digitize the the audio source in the recording environment, and faithfully reproduce it in the listening environment. The specifications of digital audio place extreme demands on converter performance. Although the storage of digital signals on compact disc (CD) and digital-audio tape (DAT) are only to 16-bit accuracy, for professional recording and reproduction systems, a resolution of 20-24 bits is desirable [Cra93, Fie89] in order to accommodate the dynamic range of the signal in the recording environment and preserve this through the various mixing and equalization stages. Additionally, for the highest quality audio, the demands on the linearity of the converter are

extreme. Ideally the converter should act as a linear channel with all noise elements uncorrelated and therefore appearing as an additive random residue [Haw90].

Due to the demands on analogue component matching, it is difficult to achieve in excess of 16-bits resolution and linearity using the well known Nyquist sampled A-D conversion techniques such as successive approximation, dual-slope integration, flash conversion [Dar90], and D-A conversion techniques such as binary weighted summation [Fie89]. The difficulties associated with these techniques stem from the fact that the accuracy of the quantization must be high in every converted sample. A class of converters in which the accuracy requirement is reduced, uses feedback around the quantizer to ensure that quantization errors in each converted sample are post-compensated by errors in subsequent samples over a specified information bandwidth. This principle is termed noise-shaping, or error spectral-shaping, and was invented by Cutler [Cut60] in 1960 and formalized by Spang and Schultheiss [Spa62] in 1962. The noise-shaping principle can also be considered in the frequency domain, where the effect of noise shaping is to reduce quantization noise power in the information band, at the expense of increased noise power out-of-band. By information theory, the capacity of the channel cannot be increased using noise shaping [Ger89]; however a higher sampling rate than the Nyquist rate can be used to ensure that there is 'excess' bandwidth in which to locate the quantization noise (figure 1.1). This concept is called oversampling. The use of noise-shaping and oversampling allows a lower quantizer resolution than would otherwise be required and in this way noise shaping trades off quantizer resolution against operating speed. This principle was first use for audio D-A converters in [Pla82] to enable 16-bit resolution to be obtained by a 14-bit D-A converter operating with four times oversampling.

A conversion strategy which utilises this trade-off to its limit is the Sigma-Delta ($\Sigma\Delta$) modulator A-D or D-A converter. $\Sigma\Delta$ modulation uses noise shaping in conjunction with a one bit quantizer to code an input signal into a high rate single bit PCM signal which observes high channel capacity over a narrow bandwidth. The key advantage of using a one-bit quantizer is that there are only two quantization levels, therefore the quantizer exhibits perfect differential nonlinearity. $\Sigma\Delta$ modulation was first described by Inose, Yasuda and Murakami [Ino63] in 1962 for the transmission of video signals. The modulator was introduced as a modified form of delta modulation [Ste75] to allow DC coding, whilst avoiding the well-known problems of slope overload distortion, and improving immunity to channel errors.

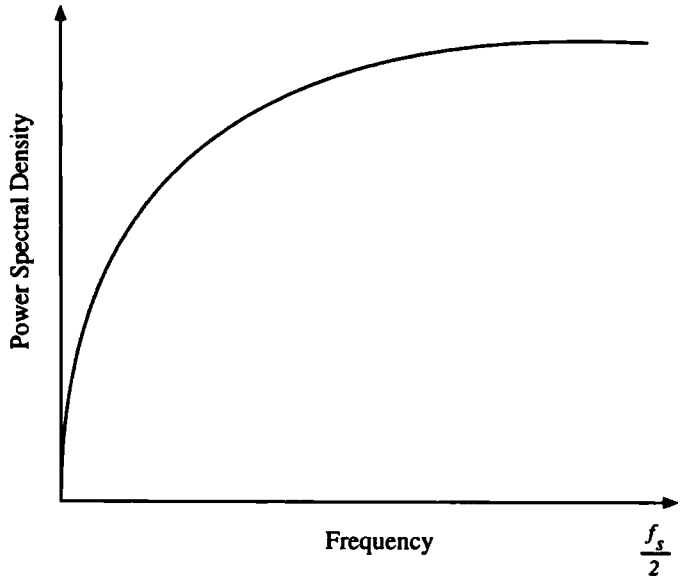


Figure 1.1: Example Noise-shaping spectrum

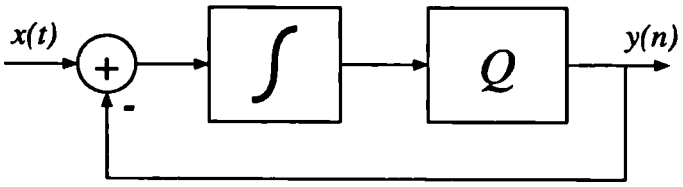


Figure 1.2: First order analogue-input ΣΔ modulator

The structure of the most basic ΣΔ modulator is shown in figure 1.2. This has a single integrator in the feedforward path and is termed a first order modulator. The key developments in ΣΔ modulation included increasing the loop to second order [Can85] and higher [Lee87, Cha90], the development of decimation structures [Can76, Can86, Chu84, Mok94], multi-stage modulators [Mat87, Rib91] and extension of the technique to bandpass modulators [Gou94, Al-95].

The majority of research papers have concentrated on the ΣΔ A-D converter, which consists of an anti-alias filter, an analogue-input ΣΔ modulator, typically utilising a switched capacitor loop-filter, followed by a decimator to convert the one bit signal, otherwise known as a bitstream, into a Nyquist rate multibit PCM signal (figure 1.3). It is also possible to configure the modulator as a D-A converter, comprising an interpolation filter to increase the sampling rate of the PCM input signal, a digital ΣΔ modulator, to convert the PCM signal to a bitstream, followed by a sample-and-hold and analogue low pass filter, which demodulates the bitstream

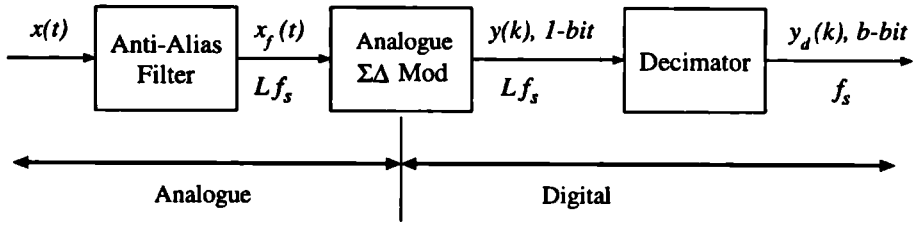


Figure 1.3: ΣΔ modulator A-D Converter

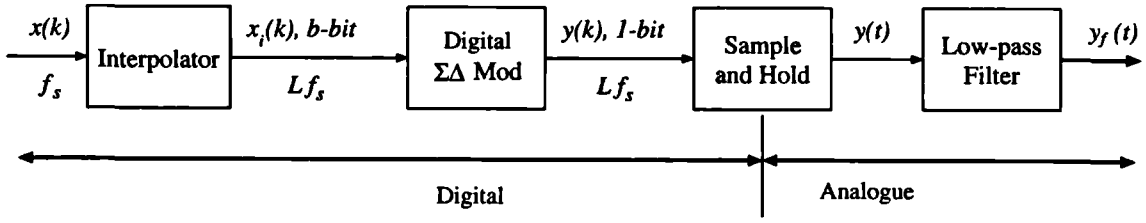


Figure 1.4: ΣΔ modulator D-A Converter

(figure 1.4). An alternative configuration implements the sample and hold as a high voltage power switch and the low-pass filter as a passive L-C network, thereby allowing power digital to analogue conversion (figure 1.5). The concept was proposed by Sandler [San83] in 1983 for audio converters, using pulse-width modulation (PWM) to produce a one-bit signal from a PCM signal. ΣΔ modulation can also be used to generate a one-bit signal. The advantage of power D-A converters in audio systems is that the digital signal is converted to analogue directly across the loudspeaker and therefore the analogue power amplifier can be eliminated from the reproduction chain. Furthermore, the use of switching (class D) amplification can lead to power efficiencies in excess of ninety percent [Ped94]. Power D-A applications also arise in control systems where precise motor control is required, and for portable systems, for example personal communication systems and hearing aids [Eng94, Hen96], where power consumption must be kept to an absolute minimum. The subject of ΣΔ modulation power D-A converters will be returned to in section 2.6.

Recently ΣΔ modulation has achieved immense popularity. In the case of ΣΔ A-D converters, the loop filter can be implemented as a switched capacitor filter allowing implementation in digital VLSI ASIC technology [Can92]. This is extremely desirable, as it allows the implementation of single-chip DSP systems with analogue interfaces, allowing a compact hardware realization. Furthermore, ΣΔ modulators are used increasingly as the fundamental building block in one-bit signal processing

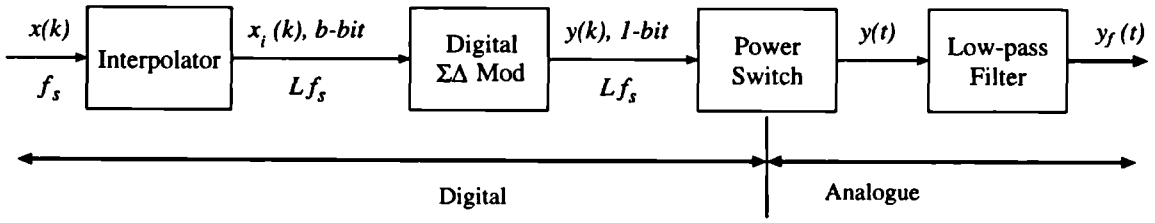


Figure 1.5: ΣΔ modulator power D-A Converter

systems, where linear filtering or other algorithmic functions are performed directly on the bitstream, without prior decimation [Ker96]. The advantage of this technique is that multipliers are no longer required, resulting in considerable hardware savings. This has lead to the use of ΣΔ modulators in applications such as function generators for on-chip testing and FM signal generation [Vei96]. The ΣΔ bitstream format is also under consideration for the mastering and archiving of audio recordings [Hi-96]. This application increases the requirement for high accuracy A-D conversion.

Despite their popularity, ΣΔ modulators have inherent problems. The modulator consists of a severe nonlinearity within a feedback loop, resulting in the occurrence of limit cycle oscillations. The occurrence of idle tones was reported in an early paper by Candy [Can76] and has been the subject of considerable research interest. Limit cycles are intrinsically linked to the ability of the modulator to code an input signal. However, due to these oscillations, for certain low-level input signals the quantization noise spectrum may become completely or partially tonal. Under this condition the modulator is said to be idling and the resulting tones are termed idle tones. Idle tones are most commonly attenuated by the addition of a noise source to the quantizer input, called dither. Higher-level oscillations may also occur, sometimes referred to as saturation limit cycles [Mot93], in which the quantizer input signal oscillates at low frequency and becomes excessively large or unbounded. These oscillations may be self sustaining for modulator orders greater than two, even when the input signal is removed [Cha90]. Under this condition the modulator does not correctly code the input signal and is said to be unstable. Instability is especially a problem with higher order modulators and care must be taken in design to avoid it.

As will be discussed in chapter 2, modulator instability limits the achievable dynamic range because a tradeoff exists between baseband noise attenuation and the maximum stable input signal. This has lead to the development of architectures which fundamentally modify the operation of the modulator in an attempt

to enhance its dynamic range. Examples of these are multi-stage $\Sigma\Delta$ architectures [Mat87], where a high order modulator is generated by the cascade of stable lower order modulators, and adaptive modulators, where either the quantizer or loop filter is adapted in order to increase dynamic range [Yu92].

1.0.1 Thesis Overview and Contributions

The research presented in this dissertation is concerned chiefly with enhancing the performance of $\Sigma\Delta$ modulators using a new class of algorithms, which operate by selective inversion of the state of the quantizer. The technique is termed *bit-flipping.* The research presented in chapters 4 to 6 are based upon the author's publications [Mag95d, Mag95a, Mag95f, Mag95e, Mag95b, Mag95c, Mag96b, Mag96a] related to bit-flipping.

Chapter 2 forms the main background theory for subsequent chapters. The simplified additive white quantization noise model is introduced and its limitations are outlined. The technique of quasi-linear modelling is introduced to address some of these limitations. The non-linear behaviour of the modulator is discussed and it is shown how this relates to non-ideal converter performance. Techniques used for linearisation are reviewed. The next section discusses $\Sigma\Delta$ architectures which adapt the operation of the modulator. The application of $\Sigma\Delta$ modulation to power digital-to-analogue conversion is discussed and alternative techniques based upon pulse-width modulation (PWM) are reviewed. Finally, the technique of bit-flipping is introduced, which forms a central theme in the chapters 4 to 6, and its potential advantages are summarised.

Chapter 3 reviews and develops the technique of quasi-linear modelling, used to predict modulator behaviour. Two methods of evaluating the quasi-linear parameters are discussed. The first evaluates the parameters with knowledge only of the noise transfer function and a probability density function (PDF) assumption. The second uses a simulation approach. The main contributions of this chapter are a method of evaluating the quantizer gain using the solution of a set of nonlinear equations, and a new interpretation of modulator stability, based upon this. The methods are developed in chapter 4 for the analysis of dithered and bit-flipping modulators.

Chapter 4 is devoted to the linearisation of $\Sigma\Delta$ modulators using bit-flipping. A new interpretation of the effect of dither on the operation of the quantizer is used

to show how bit-flipping can linearise the modulator. A bit-flipping algorithm is then developed which emulates dither. It is shown that bit-flipping can increase the conversion linearity without the use of a random element. This technique is termed deterministic bit-flipping (DBF). By developing the quasi-linear model to deal with modified quantizers, the noise and stability of DBF is compared to rectangular PDF dither and a theoretical relationship between the two techniques is identified. This is followed by simulation results, which compare in detail the performance of DBF to different types of dither. It is concluded that under certain conditions, DBF offers lower noise penalties, and is well suited to A-D converter implementation.

In chapter 5, it is shown how bit flipping can be used to modify the spectrum of the quantization noise so that improved baseband noise shaping is obtained. It is shown how the bit-flipping makes the modulator adaptive, which leads to an enhancement in dynamic range. By the development of an equivalent model, it is shown how bit-flipping may conceptually increase either the noise performance or stability of the modulator. Simulation results are presented which highlight key design issues of bit-flipping modulators and it is shown that, in comparison to fixed modulators, improvements in dynamic range can be achieved, though care must be taken to avoid internal overload.

Chapter 6 concentrates on the application of bit-flipping to $\Sigma\Delta$ power digital-to-analogue converters. The bit-flipping technique is used to reducing the pulse-repetition rate of the bitstream in order to lower the power dissipation and improve the error tolerance of the subsequent power switching stage. A number of algorithms are presented which aim to optimise the stability and noise performance of the bit-flipping modulator. A key advantage of the proposed technique is that improved linearity and reduced master clock speeds are possible, in comparison to more conventional PWM-based converters. The reduced clock speed allows straightforward implementation using ASIC technology.

Finally chapter 7 summarises the results and presents suggestions on open areas of future research.

Chapter 2

Background on Sigma-Delta Modulation

2.1 Introduction

This chapter presents the aspects of $\Sigma\Delta$ modulation and related techniques which are relevant to work in subsequent chapters. The aim is to provide a general understanding of $\Sigma\Delta$ modulation, with emphasis on modulator design and performance. The work briefly reviews simplified modelling and design methods, nonlinear behaviour and idle tones, linearisation techniques, adaptive modulators and power conversion techniques. Finally the central concept of the bit-flipping modulator is introduced.

2.2 Modelling Techniques

Studies into the behaviour of $\Sigma\Delta$ modulation have generally been approached from one of two separate perspectives.

In the first, assumptions and approximations are made which allow the system analysis to be approached from standard linear or quasi-linear techniques. This approach aids the intuitive understanding of the concepts of noise-shaping and one-bit coding, whilst approximately relating modulator parameters to performance. The techniques developed are especially useful in guiding the design process, which may then be completed iteratively by means of computer simulation.

In the second, the modulator is recognised as a nonlinear system and analysed

using nonlinear methods. This approach encompasses the methods of exact analysis, and nonlinear dynamical analysis. Exact analysis methods include papers such as [Gra89], [Fri88] and [He90], in which the exact quantization error spectrum is evaluated. These methods have been applied for both DC and periodic inputs [Del92] but have lacked sufficient generality to be applied to modulators with third or higher order loop filters

Techniques which have recently gained considerable popularity model the $\Sigma\Delta$ modulator as a nonlinear dynamic system, and examples include [Ris94b, Fee95, Dav96]. These methods are normally formulated towards the rigorous characterisation of modulator stability and signal bounds, by means of identifying regions in the modulator state-space which correspond to low-level and saturation limit cycles. They have been applied in identifying stability properties of both chaotic and non-chaotic modulators with first and second order loop filters [Ris94b]. The methods are generally not formulated towards evaluation of noise performance, however.

Due to the difficulty in analysing the nonlinearity of the one-bit quantizer, these methods tend to lack generality and at the time of writing no published studies have found solutions to generalised modulators of third or higher orders. Furthermore, in this dissertation, we are mainly concerned with the noise and stability properties of general modulators with third or higher order loop filters, therefore the use of linear and quasi-linear techniques must be relied upon, and the necessary assumptions and approximations must be tolerated. Nonlinear analysis is useful, however, in explaining the idle tone phenomenon, and the subject will be returned to briefly in section 2.4.

2.2.1 Linear Modelling

In this section linear modelling is reviewed and some of assumptions and approximations are examined. This is followed by a discussion of quasi-linear modelling, which forms the background to the work in chapter 3.

The discrete-time model of a $\Sigma\Delta$ modulator is shown in figure 2.1. Many structures have appeared which differ slightly from this form, for example the multiple feedback structure [Iso91] and the noise-shaping structure [Sti88b]; however these can usually be reworked into the form of figure 2.1 by the inclusion of a signal pre-filtering network [Sch90]. The discrete-time model is also relevant to $\Sigma\Delta$ A-D converters employing switched capacitor loop filters, and with appropriate transfor-

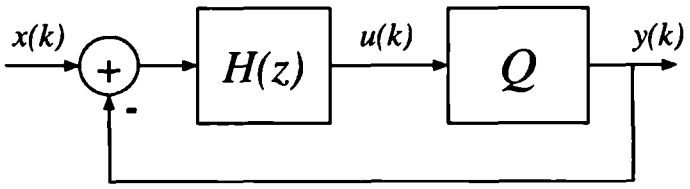


Figure 2.1: Discrete-time Model of ΣΔ Modulator

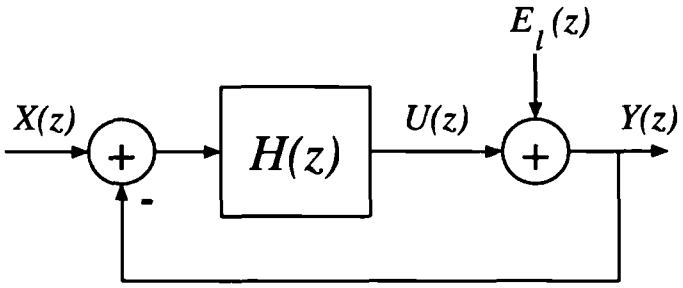


Figure 2.2: Discrete-time Model with Additive Error

mations, to modulators employing continuous time loop filters [Cha90]. Without making any approximations, the quantizer may be replaced by an additive error source $E_l(z)$, defined as $E_l(z) = Y(z) - U(z)$ (figure 2.2) and this enables the modulator to be modelled as a dual-input linear system, as shown in figure 2.3. It is important to realize that $E_l(z)$ is internally generated by the system; however the concept of linear modelling is that $E_l(z)$ may be approximated by an independent noise source of known properties. This concept was used originally for noise shaped systems in [Ino63, Spa62] and has since appeared throughout the literature. By linear systems theory the z-domain equation for the modulator output is given by:

$$Y(z) = X(z)STF(z) + E_l(z)NTF(z) \tag{2.1}$$

where

$$STF(z) = \frac{H(z)}{1 + H(z)} \tag{2.2}$$

$$NTF(z) = \frac{1}{1 + H(z)} \tag{2.3}$$

The expression $STF(z)$ is termed the *signal transfer function* (STF) and this defines the relationship between the input signal $X(z)$ and the signal component of the output signal $Y(z)$. The expression $NTF(z)$ is the *noise transfer function*

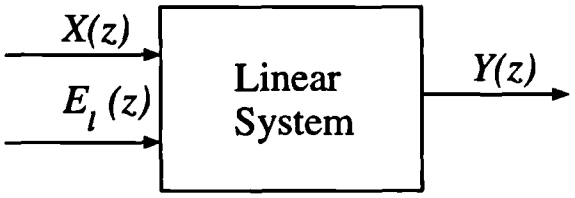


Figure 2.3: Dual-input Linear Model

(NTF) and this defines the relationship between the error signal $E_l(z)$ and the error component of the output signal.

The expression 2.1 demonstrates that if $E_l(z)$ were an independent input with a known spectrum, superposition would apply, allowing the noise spectrum in the output signal to be defined by appropriate NTF design. In this dissertation we consider applications where $NTF(z)$ is high-pass, therefore the quantization noise is attenuated in the baseband (at the expense of increased noise out of band). The independence of $E_l(z)$ is of course crucial to these relationships, since any correlation between $X(z)$ and $E_l(z)$ will cause the STF and NTF to become interlinked and superposition will no longer apply.

The popular assumptions on the characteristics of $E_l(z)$ which appear in the literature (see for example [Ino63, Can86, Agr83, Cha90]) are that $E_l(z)$ is an independent additive white noise source with variance $(2\Delta)^2/12$, where $\pm\Delta$ are the quantizer output levels. This approximation originates from the work of Bennet [Ben48], in which conditions are stated for a quantization error which is approximately uncorrelated (i.e. a white noise source) and uncorrelated with the quantizer input $U(z)$:

1. The number of quantizer output points is large.
2. Successive input samples have small correlation.
3. The output points are very close to the midpoints of the set of inputs which yield these points.

Gray [Gra87] discusses how the quantizer approximation is poor for the single bit quantizer of the $\Sigma\Delta$ modulator. Firstly, and most seriously, the quantizer cannot be considered to have a large number of levels. Secondly, the quantizer is within a feedback loop, and so successive inputs to the quantizer cannot be considered to be uncorrelated. Gray states that the validity of the third condition is 'not

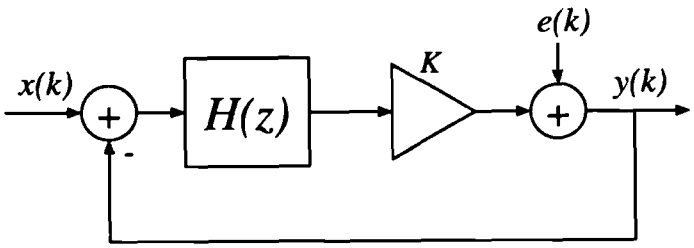


Figure 2.4: Quasi-Linear Model of ΣΔ Modulator

known', however the condition implies that the quantizer input must not overload, i.e. the quantizer input must be bounded within the two output levels. The concept of quantizer overload is ambiguous in the case of a one-bit quantizer because the quantizer responds only to the sign of its input, therefore a positive scaling of $U(z)$ by a gain term G does not influence the operation of the modulator, assuming that there is no hardware limitation on maximum signal levels. This property is termed *scaling invariance* [Ris94b]. Of course, such a scaling in the linear model would affect the predicted noise power, and this demonstrates further the inadequacy of the model.

2.2.2 Quasi-linear Modelling

Ardalan and Paulos [Ard87] develop an improved method for modelling the quantizer, called quasi-linear modelling, in which the AC component of the quantizer error is modelled as a linear gain term K_n followed by an additive white noise source $e(k)$ of variance σ_e^2 (figure 2.4). The model originates from the statistical describing function method of Booton [Boo53] for modelling nonlinear elements in feedback control systems. Useful overviews of the describing function technique are presented by Smith [Smi66] and by Atherton [Ath82]. The values of K_n and σ_e^2 are defined according to the steady-state statistics of the quantizer input and output signals, under the assumption that the signals are quasi-stationary processes. Quasi-stationary signals have well defined first and second moments (i.e. the mean and variance), and this allows traditional correlation and spectral analysis techniques to be used [Cho91].

The values are input signal dependent and allow the model to predict an observed change in noise performance with the amplitude of the steady-state input signal. The gain term also ensures that any scaling of $H(z)$ by G is accommodated in the model

by an inverse scaling $K_n = 1/G$.

A simple way of defining K_n has been described in [Ada91]: simply that K_n is the ratio of the mean of the quantizer output and input. In chapter 3, a more mathematically rigorous method will be developed, in which the values of K_n and σ_e^2 are defined in such a way as to ensure that the quantizer error is uncorrelated from its input. This allows $e(k)$ to be modelled as an external noise source. Although no rigorous proofs appear in the literature, empirical results have suggested that the error signal in the quasi-linear model approximates more closely to a white noise source than in the linear model (refer to appendix B.1).

The inclusion of the quasi-linear gain K_n modifies the STF and NTF. We are mainly concerned with the NTF, which becomes:

$$NTF_K(z) = \frac{1}{1 + H(z)K_n} \tag{2.4}$$

Using the white noise assumption, the baseband noise power can be expressed in terms of the NTF and the variance of the quantization noise σ_e^2:

$$P_b = \frac{\sigma_e^2}{\pi} \int_0^{\pi/L} |NTF_K(\theta)|^2 d\theta \tag{2.5}$$

where L is the oversampling ratio, defined as the ratio of the oversampling frequency f_o and the Nyquist rate sampling frequency f_s:

$$L = \frac{f_o}{f_s} \tag{2.6}$$

and θ is the normalised frequency in radians:

$$\theta = \frac{2\pi f}{f_o} \tag{2.7}$$

for a frequency f in Hz.

Since K_n and σ_e^2 depend on the steady state input (assuming stationary signals) the baseband noise power P_b is constant for a steady-state input but may change with input signal amplitude. This allows the model to predict observed changes in noise power with input level. These concepts will be developed in chapter 3.

2.2.3 Stability of Sigma-Delta Modulation

The stability of the $\Sigma\Delta$ modulator has been the focus of considerable research interest, primarily because modulators of second and higher order are only conditionally

stable [Cha90]. By modulator *order* we mean the order of the NTF. In the literature it is common to use the term 'low-order' to describe a first and second order modulator, and 'high-order' to describe a third or higher order modulator.

In this section stability is discussed with reference to the quasi-linear model. We begin with a definition of stability, based upon the coding accuracy of a $\Sigma\Delta$ modulator.

The traditional definitions of stability are:

1. Bounded Input Bounded Output (BIBO)

2. Bounded Input Bounded State (BIBS)

Both of these definitions are inadequate for $\Sigma\Delta$ modulation. The first definition is inappropriate because the modulator output is binary and therefore cannot become unbounded. The second is also inappropriate because, for a modulator with $H(z)$ poles within the unit circle, the internal states are always bounded [Sch93]. A more appropriate definition of stability is simply that the modulator is stable if it accurately codes the input signal. This definition can be applied most easily to a modulator with a DC input. The case of a DC input is used frequently in analysis (see for example [Ard87, Ris94b]) to approximate a time-varying baseband input, the rationale being that if the oversampling ratio is high enough, a baseband limited signal will change very slowly relative to the oversampling frequency, and may be approximated by a DC signal [Hei91]. For a DC input, the steady state response of the modulator from equation 2.2 is:

$$E\{y(k)\} = E\{x(k)\}\frac{H(1)}{1+H(1)} \tag{2.8}$$

where $E\{\}$ is the expectation operator i.e. it represents the statistical mean.

For a baseband modulator $H(1) >> 1$, hence $E\{y(k)\} \approx E\{x(k)\}$. Therefore for DC inputs, the stability definition can be expressed as follows:

Property 2.1 (DC Coding Property) *For a stable modulator with DC input* $x(k) = m_x$, $E\{y(k)\} \approx E\{x(k)\}$

Conversely, for an unstable modulator, $E\{y(k)\} \neq E\{x(k)\}$. This property will be demonstrated in section 2.3.3.

Insight into whether the DC coding property may be observed for a particular modulator and input can be gained from the following property.

Property 2.2 (Constant Output Power Property) *For a binary quantizer with levels* $\pm\Delta$*, the output power is a constant* Δ^2*.*

This property is interesting because it means that for a one-bit quantizer, the output power is a constant which is shared between signal and noise power. For an output signal with DC component $E\{y(k)\} = m_y$ and noise variance σ_q^2, the Constant Output Power Property may be applied to yield the following equality:

$$m_y^2 + \sigma_q^2 = \Delta^2 \tag{2.9}$$

By the DC coding property, this equality may be rewritten as:

$$m_x^2 + \sigma_q^2 \approx \Delta^2 \tag{2.10}$$

which is true for any stable modulator. Incidentally, the constant output power does not fit into the framework of the linear model because there is no mechanism by which an increase in signal level can be balanced by a reduction in noise power; however in the quasi-linear model the values of K_n and σ_e^2 are able to adjust in order to maintain the equality.

The value of σ_q^2 is obtained as

$$\sigma_q^2 = \frac{1}{\pi}\int_0^{\pi} |E(\theta)|^2 |NTF_K(\theta)|^2 d\theta \tag{2.11}$$

This expression can be simplified by applying the white noise assumption.

$$\sigma_q^2 = \sigma_e^2 P_n(K_n) \tag{2.12}$$

where $P_n(K_n)$ is the power gain of the NTF, which is a measure of the power amplification of the quantizer noise obtained at the modulator output.

$$P_n(K_n) = \frac{1}{\pi}\int_0^{\pi} |NTF_K(\theta)|^2 d\theta \tag{2.13}$$

By Parseval's Relation [Pro92], this may also be expressed in the time domain from the NTF impulse response as:

$$P_n(K_n) = \sum_{k=0}^{\infty} |ntf(k)|^2 \tag{2.14}$$

Combining equations 2.10 and 2.12, for a stable modulator:

$$m_x^2 + \sigma_e^2 P_n(K_n) \approx \Delta^2 \tag{2.15}$$

This equation is very revealing with regards to the stability of the modulator. It will be shown in chapter 3 that for high order modulators, $P_n(K_n)$ has a minimum which is greater than zero, and this places an upper limit on the value of m_x. Increasing the input level beyond this value causes a breakdown in the DC coding property and the modulator becomes unstable.

The maximum input signal level just before the coding break-down is called the Maximum Stable Amplitude (MSA). The input power at which the coding breaks down depends upon the noise power amplification i.e. the power gain of the NTF.

2.3 Noise-Transfer Function Design

The behaviour of a modulator for a given input $X(z)$ is determined entirely by the NTF coefficients, since these are the only free modulator parameters, therefore the NTF completely characterises the modulator and its design is of fundamental importance. It will be seen that $NTF(z)$ defines not only the noise properties of the modulator, but also its stability and linearity. For baseband coding, for example in audio applications, $NTF(z)$ is designed as a high-pass function so that the quantization noise is attenuated in the baseband, at the expense of increased quantization noise out-of-band. The out-of-band noise can be later filtered by decimation in the case of an A-D converter, or by analogue low-pass filtering, in the case of a D-A converter. [1]

Design methodologies for $\Sigma\Delta$ NTFs have been discussed in several places, including [Tew78, Cha90, Agr83, Sch90, Ada91, Ris94b].

Two important principles apply when designing general noise-shaping coders (either with single bit or multibit quantizers): the NTF Scaling Constraint and the Noise Shaping Theorem. In these, the value of the parameter K_n is defined by the modulator in operation (refer to chapter 3), rather than being specified in the design.

[1] Note that the noise spectrum at the D-A converter output is influenced also by the choice of 1-bit conversion. A sample-and-hold D-A converter has a sinc function with zeros at multiples of the oversampling frequency, whereas an impulse D-A converter has a flat response [Ris94a]

Theorem 2.1 (NTF Scaling Constraint [Ada91]) *For implementability the transfer function $NTF_K(z)$ must be scaled such that first term in its impulse response is unity i.e. $ntf_K(0) = 1$*

The scaling factor, sometimes termed the *necessary* scaling [Ris94b], ensures that there is a unit delay in $H(z)$, thus ensuring that the modulator can be implemented. The scaling ensures that the NTF power gain is always greater than or equal to unity (equation 2.14), therefore the quantization noise cannot be attenuated in one band without an increase in noise power in another band. This is expressed in more precise terms in Gerzon and Craven's Noise Shaping Theorem [Ger89]. This theorem is restated below, within the framework of quasi-linear modelling.

Theorem 2.2 (Noise Shaping Theorem[Ger89]) *A transfer function $NTF_K(z)$ scaled such that $ntf_K(0) = 1$, satisfies the relationship:*

$$\int_0^{\pi} log|NTF_K(\theta)|d\theta \geq 0 \tag{2.16}$$

Equality is obtained if and only if $NTF_K(z)$ is minimum phase.

The Noise Shaping Theorem can be interpreted by plotting $|NTF_K(\theta)|$ on a log-amplitude linear-frequency scale (figure 2.5). The theorem states that for a NTF with necessary scaling, the area above the 0 dB reference line (A) must be equal to or greater that the area below the line (B). The reference line represents the spectrum of the quantization error $E(z)$, measured in the loop. No assumptions are made about the characteristics of the quantizer, nor its error, for instance a non-white quantization error simply modifies the shape of the reference line.

From an information theory perspective, area A represents the frequency band where the information capacity is reduced, and conversely area B represents the band where the information capacity is increased. The noise shaping theory dictates the scaling required to ensure that the overall information capacity of the noise-shaping channel is not increased and therefore Shannon's information theory is satisfied. For a minimum phase filter (i.e. all zeros of $NTF_K(z)$ within the unit circle), the necessary scaling is precisely that which results in the areas above and below the 0 dB line becoming equal. For the case of a non-minimum phase filter, the inequality in equation 2.2 implies that the $A > B$ and this will be established by the necessary scaling. Gerzon [Ger89] states that the capacity of the channel is preserved using minimum phase noise shaping filters, therefore these types of filters are optimal.

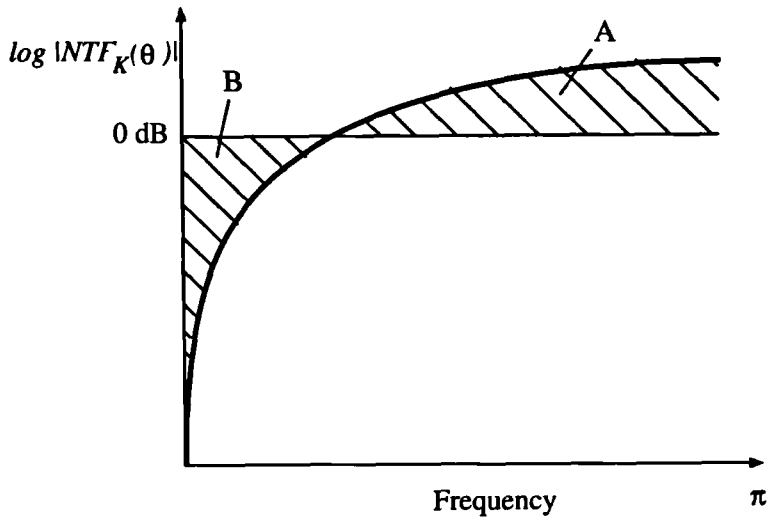


Figure 2.5: Error Spectrum representation of Noise Shaping Theorem

2.3.1 Noise Transfer Function Design and Tradeoffs

We will now show how the NTF Scaling Constraint and Noise Shaping Theorem may be used to design NTFs using an example filter class, which will be used as the basis for all modulator NTFs in the following chapters. This subject has also been discussed in [Cha90, Ada91, Sch90, Ris94b].

It is important to recognise that for any NTF filter class, the value of K_n associated with the quantizer will influence the NTF frequency response. This is unfortunate from a design perspective, since K_n depends on the input level and will also change if any modifications are made to the quantizer operation, for example with bit-flipping algorithms. For convenience, the NTFs are therefore designed here with a value of $K_n = 1$. In terms of baseband attenuation, the influence of any deviation in K_n from unity can be approximated by assuming $H(z) >> 1$ in the signal band, therefore $NTF_K(z) \approx \frac{1}{H(z)K}$. This approximation is substituted into equation 2.5, giving:

$$P_b \approx \frac{\sigma_e^2}{\pi K_n^2} \int_0^{\pi/L} \frac{1}{|H(\theta)|^2} d\theta \tag{2.17}$$

Therefore the baseband noise power will be directly scaled by the value of K_n.

The NTFs designed here are based upon the infinite-impulse response (IIR) Butterworth high pass filter class, because its flat frequency response above the cutoff frequency ensures that the out-of-band quantization noise is evenly distributed at high frequencies (assuming a white quantization noise). Furthermore, the STF has

BIBL. LAUSANNE EPFL

Order N	Zero Frequencies
1	0
2	$\pm 1/\sqrt{3}$
3	$0, \pm\sqrt{3/5}$
4	$\pm\sqrt{3/7 \pm \sqrt{(3/7)^2 - 3/35}}$
5	$0, \pm\sqrt{5/9 \pm \sqrt{(5/9)^2 - 5/21}}$
6	$\pm 0.23862, \pm 0.66121, \pm 0.93247$
7	$0, \pm 0.40585, \pm 0.74153, \pm 0.94911$
8	$\pm 0.18343, \pm 0.52553, \pm 0.79667, \pm 0.96029$

Table 2.1: NTF Optimal Zero Frequencies normalised to θ_B

Butterworth poles, resulting in a desirable phase and magnitude response [Sch90]. The baseband response is tailored by the distribution of zeros on the unit circle, across the baseband. The zero locations have been defined exactly as in [Sch90], as these locations have been shown to minimise the baseband noise power. The zero frequencies are given in table 2.1, normalised to θ_B, where $\theta_B = 2\pi/L$

Using this class of filters, for a given oversampling ratio L and order N, the frequency response is defined by a single free parameter - the normalised cutoff frequency $f_{nc} = f_c/Lf_s$, where f_c is the cutoff frequency in Hz.

Consider the example of three filters with $L = 64$ and normalised cutoff frequencies of 1/40, 1/16 and 1/10. The filter with the highest cutoff frequency has the greatest baseband attenuation because of the greater separation between poles and zeros. All three filters have unity gain at $Lf_s/2$. Applying the NTF scaling constraint results in all three frequency responses being scaled 'upwards' (figure 2.6) in accordance with the Noise Shaping Theorem. The scaling is greatest for the filter with the highest cutoff frequency because it has the highest baseband attenuation and the widest transition region. A wide transition region is 'expensive' in terms of noise shaping because the extra area below the 0 dB line must be compensated for by an increase in area above the line (therefore the 'upwards' scaling must be greater).

The area (A) above the 0 dB line is important in terms of the stability of the modulator. Equation 2.13 relates this area to the power gain of the NTF (making the assumption that the baseband attenuation is high and makes little contribution

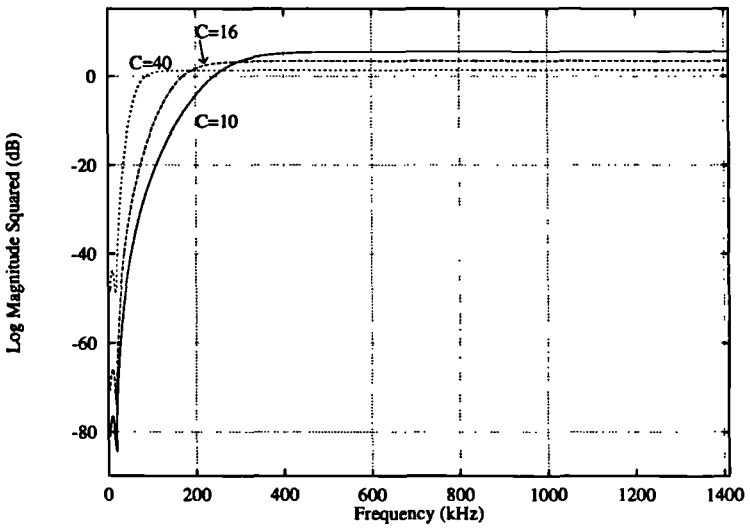


Figure 2.6: Frequency responses of three Butterworth filters after scaling, $C = f_c/f$

to the integral) and shows that increasing the area results in an increased power gain. There is therefore a fundamental tradeoff between baseband attenuation and stability. For fixed modulators, this tradeoff can only be improved by increasing the loop order, or the oversampling ratio. Increasing the loop order narrows the transition band for the same high frequency gain, and the area 'gained' can be used to increase the baseband area and its associated attenuation. Alternatively, the stability can be increased by reducing the the high frequency gain for the same baseband attenuation. Increasing the loop order is generally undesirable because the hardware cost increases and sharper filtering is required in the decimation or reconstruction filters. Another method of improving the tradeoff is to increase the oversampling ratio. At the expense of an increase in operating speed, the quantization noise is spread over a greater bandwidth, and both the baseband and high frequency noise is reduced. Improving the noise-stability tradeoff without increasing the loop order or oversampling ratio is the subject of the research presented in chapter 5.

2.3.2 Finite-Impulse Response (FIR) Noise Transfer Functions

The Noise Shaping Theorem also explains why a popular class of noise-shaping filters yield unstable operation for higher than second order modulation. Tewksbury and Halloc [Tew78] define a general class of N^{th} order FIR NTFs with transfer function:

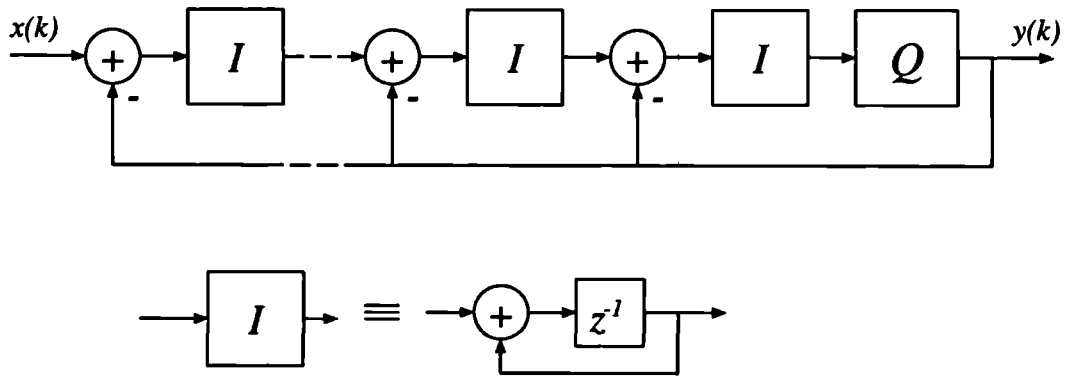


Figure 2.7: Discrete-time Structure of N^{th} order Interpolative Coder.

(we again assume $K_n = 1$)

$$NTF(z) = (1 - z^{-1})^N \tag{2.18}$$

These filters are optimal in the sense of minimising the baseband noise power in a manner which is independent of oversampling ratio. With $N = 1$ and $N = 2$, the filters correspond to the first order modulator described by Inose et al [Ino63] and the second order modulator described by Candy [Can85]. The general structure for an N^{th} order modulator is described by Ritchie (reference 13 of [Tew78]) and this is shown in figure 2.7. The disadvantage of these structures is that the NTF has a very wide transition band leading to a high power gain by the Noise Shaping Theorem. This has led some practitioners to falsely believe that all modulators higher than second order are impossible to stabilise [Can85].

2.3.3 Demonstration of DC Coding Property

In this section we briefly consider the DC coding accuracy of three modulators in order to demonstrate typical stability properties of different modulator NTFs. The three modulator NTFs are:

- A: First order Tewksbury $NTF(z) = 1 - z^{-1}$.
- B: Second order Tewksbury $NTF(z) = (1 - z^{-1})^2$.
- C: Third order Butterworth with optimised zeros (table 2.1) and cutoff frequency yielding a power gain $P_n = 3.5\ dB$ after necessary scaling.

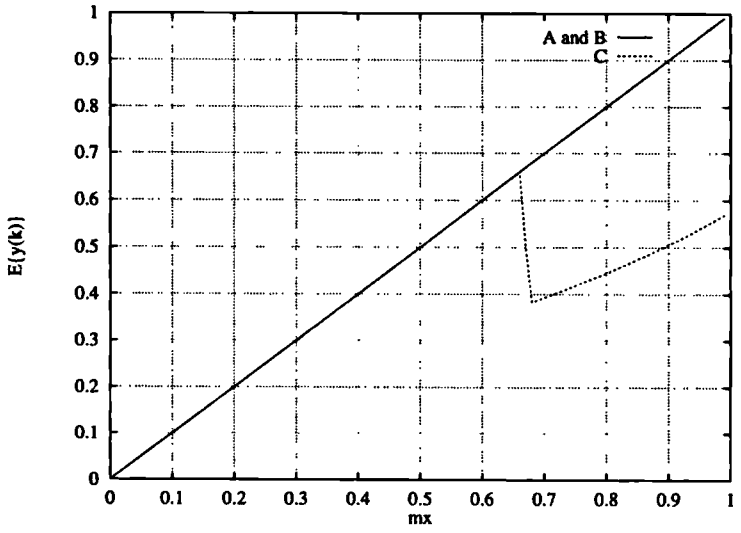


Figure 2.8: Coding Accuracy for Examples A, B, C

The modulators have been simulated for $M = 100,000$ samples with a DC input m_x in the range $m_x = 0 \rightarrow 1$. The first 1000 samples have been discarded to allow the modulator to settle from the initial transient (refer to appendix A.1 for details of the simulation scheme.) The mean output level has been estimated by:

$$E\{y(k)\} = \frac{1}{M} \sum_{0}^{M-1} y(k) \tag{2.19}$$

In figure 2.8 $E\{y(k)\}$ is plotted against m_x for the three modulators. The first and second order modulators follow the same characteristics, with the DC coding $E\{y(k)\} \approx m_x$ maintained up to full scale. The coding of the third order modulator fails at $m_x = 0.66$. At higher input levels, the quantizer input signal grows to very high signal level, demonstrating instability. These results shows how the DC Coding Property (property 2.1) can be used to identify modulator stability and also indicate the conditional stability of higher order modulators.

2.4 Nonlinear Behaviour in Sigma-Delta Modulators

Although the quasi-linear model can be used to predict the noise performance and to some degree the stability of $\Sigma\Delta$ modulators, there are certain aspects of modulator behaviour which cannot be explained by linear or quasi-linear theory. One particular

aspect of modulator performance which is typical of a nonlinear feedback system is the presence of idle tones in the noise spectrum. The occurrence of tones is one of the most critical problems because the linearity of the conversion can be seriously degraded. For audio conversion, the predominance of tones at low input levels is especially problematic because perceptual signal masking effects are small [Nor95]. The ear is also particularly adept at distinguishing periodic components in the presence of noise [Fie89], therefore idle tones can be subjectively very disturbing [Sti88a].

The occurrence of idle tones was reported in an early paper by Candy [Can76], who observed noisy modes of operation for small sinewave inputs in the presence of a rational DC bias. Crucially, idle tones have subsequently been reported in varying degrees in all practical modulator orders [Nor93, Ris94a] and also in higher order delta modulators [Ada84].

The tone phenomenon has therefore been prominent in the literature, both from the perspective of performance (see for example [Sti88a, Nor93, Led3]) and analysis. In the latter case, the idle tones have been related to the occurrence of periodic patterns termed *limit cycles* [Can74] in the modulator output. This has led to the description of modulator operation using the techniques of nonlinear dynamical analysis.

2.4.1 Limit Cycles in Nonlinear Dynamical Systems

A N^{th} order $\Sigma\Delta$ modulator can be described as a nonlinear discrete time dynamical system which can be analysed by observing trajectories in the N dimensional state-space. [Ris95]. In general the state-space has a large number of periodic trajectories called limit cycles. The behaviour of the trajectories depends on the location of the poles in the loop filter $H(z)$, which correspond to the zeros of the NTF. If all the poles of $H(z)$ are inside the unit circle, the limit cycles have attracting regions in the state space, and all sufficiently small perturbations about the periodic trajectories tend towards zero [Mot93] . These are termed asymptotically stable limit cycles. Conversely if one or more of the poles of $H(z)$ are outside the unit circle, infinitesimal perturbations about the periodic trajectories result in divergence and these are termed unstable limit cycles.

It is important to realise that the stability of the limit cycles does not correspond to the stability of the modulator, since the poles of $H(z)$ correspond to the zeros of

the NTF, which do not directly affect the power gain of the modulator. Modulators with $H(z)$ poles outside the unit circle are termed *chaotic modulators*. In general, chaotic systems are unpredictable on long time scales and generate non-periodic outputs [Ris95].

In terms of the tonal behaviour of $\Sigma\Delta$ modulators, nonlinear dynamical analysis is useful because it explains the existence of periodicity in the bitstream, corresponding to the spectral occurrence of tones.

2.4.2 Identification of Tone Frequencies for DC Inputs

For a modulator with a rational DC input, a relationship exists between the tone frequencies and the input level [Led3, Ris94b]. This relationship will prove useful in chapter 4 for measuring the linearity of $\Sigma\Delta$ modulators with different dithering schemes.

Consider an output sequence of period P comprising P_1 positive bits and P_2 negative bits, such that

$$P_1 + P_2 = P \tag{2.20}$$

By the DC Coding Property 2.1, for a rational DC input m_x:

$$\frac{P_1 - P_2}{P} \approx m_x \tag{2.21}$$

Combining equations 2.20 and 2.21:

$$\frac{P_1}{P} \approx \frac{1 + m_x}{2} \tag{2.22}$$

The ratio P_1/P can thus be evaluated and the minimum period P obtained by reducing P_1/P to its lowest possible terms. The minimum period P represents the smallest period limit cycle in $y(k)$ which can represent the DC input. Obviously, for the limit cycle to exist, P_1/P must be rational, otherwise it is not possible to represent the DC level as a periodic signal. This explains why tones occur predominantly for rational inputs (see for example [Can76]).

Ledzius [Led3] has described the relationship between limit cycle period and DC input level in a more intuitive manner, by considering possible bit-patterns in the one-bit output. Firstly, the assumption is made that for zero input, a periodic pattern consisting of alternate 1's and -1's is produced (figure 2.9(a)). This pattern is the only steady-state solution for a first order modulator with zero input [Ris94b] and can also theoretically occur for higher order modulators [Nau91].

(a) 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1 1 -1

(b) 1 -1 1 1 1 -1 1 -1 1 -1 1 1 1 -1 1 -1

(c) 1 -1 1 1 -1 1 -1 1 1 -1 1 1 -1 1 -1 1

(d) 1 1 1 -1 -1 -1 -1 -1 1 1 1 -1 -1 -1 -1 -1

Figure 2.9: Representation of DC levels by Limit Cycles: a) $m_x = 0$ (8 cycles) b) $m_x = 1/4$, minimum-period pattern (two cycles) c) $m_x = 1/4$, phase-inversion pattern (two cycles) d) convolution pattern (two cycles)

By the DC Coding Property, for a small rational DC input, the DC level must be represented in the average level of the output. Here we consider the example of an input level $m_x = 1/4$. Ledzius considers two cases in which the increase in DC level can be represented. In the first case a 1 changes to a -1 in each period causing periodic {1, 1, 1} triplets in the {1, -1} pattern, figure 2.9(b). This causes a minimum period pattern and by equation 2.22, $P_1/P = 5/8$, therefore the period is eight samples. The fundamental period causes a low frequency tone to occur at a frequency $f_l = m_x L f_s/2 = L f_s/8$. In the second case, the increased DC input is represented by a regular phase inversion, figure 2.9(c), causing double '1' patterns to occur. The fundamental period is still eight samples; however, the average repetition rate of the pseudo-periodic phase inversion is now four samples, causing a dominant low frequency tone at frequency $f_l = m_x L f_s = L f_s/4$. In [Mot96], Motamed states that from empirical evidence, the {+1, -1} pattern will be distributed as evenly as possible, implying that the phase inversion pattern is more likely to occur.

In association with the low frequency tone, a high frequency tone near $L f_s/2$ is also produced [Ris94b]. The frequency of this tone can be obtained by recognising that pattern (c) is a time-domain multiplication of pattern (a) and a new pattern (d) which has a strong spectral component at $m_x L f_s/2 = f_s/8$ [Mot96]. The equivalent operation in the frequency domain is convolution, therefore in the frequency domain, the tone at $L f_s/2$ is modulated by the tone at $L f_s/8$ resulting in a high frequency tone $f_h = (1 - m_x) L f_s/2 = 3 L f_s/8$.

In summary, assuming the phase inversion pattern dominates, a rational DC input m_x will produce two dominant tones:

$$f_l = m_x L f_s \tag{2.23}$$

$$f_h = (1 - m_x)\frac{L f_s}{2} \tag{2.24}$$

A tone is generally not produced at $m_x L f_s/2$ because the presence of the pseudo-periodic phase inversion component at $m_x L f_s$ causes cancellation [Led3] of the tone at $m_x L f_s/2$.

As an example, in figure 2.10 the baseband and wideband spectra of a fourth order, 64 times oversampled modulator with a DC input of $m_x = 1/512$ and a NTF with power gain $P_n = 4.25\ dB$ are plotted. For clarity only the frequency range between 1200 kHz and $Lf_s/2$ is shown in the wideband plot. The simulation details and parameters are discussed in appendix A. As predicted by equations 2.23 and 2.24, the dominant baseband and high frequency tones falls at $f_l = 5512.5\ Hz$ and $f_h = 1408443.75\ Hz$, respectively. The tones are also combined with a shaped wide-band noise component, indicating that the limit cycles are long and complex, but retain a strong periodic component at the dominant tone frequency. In the baseband plot, additional tones are present at harmonics of the dominant baseband tone and in the wideband plot, lower sidebands of the high frequency tone are present. These sidebands are separated by the frequency of the baseband dominant tone and can be considered to be intermodulation products of the baseband harmonics and the Nyquist frequency.

The high frequency tone f_h and sidebands can also affect the linearity of the converter since any clock frequencies present around $Lf_s/2$ may cause intermodulation products to fall into the baseband [Har92].

The behaviour of the tones with increasing DC bias is also very interesting. As the DC bias increases, the period necessary to represent the input reduces, and by equations 2.23 and 2.24, the low-frequency tone increases in frequency, and the high frequency tone reduces in frequency.

It should be emphasised that these relationships are based upon a simplified analysis and do not take into account issues such as initial integrator states, which may modify the tonal composition for non-chaotic modulators [Mot96].

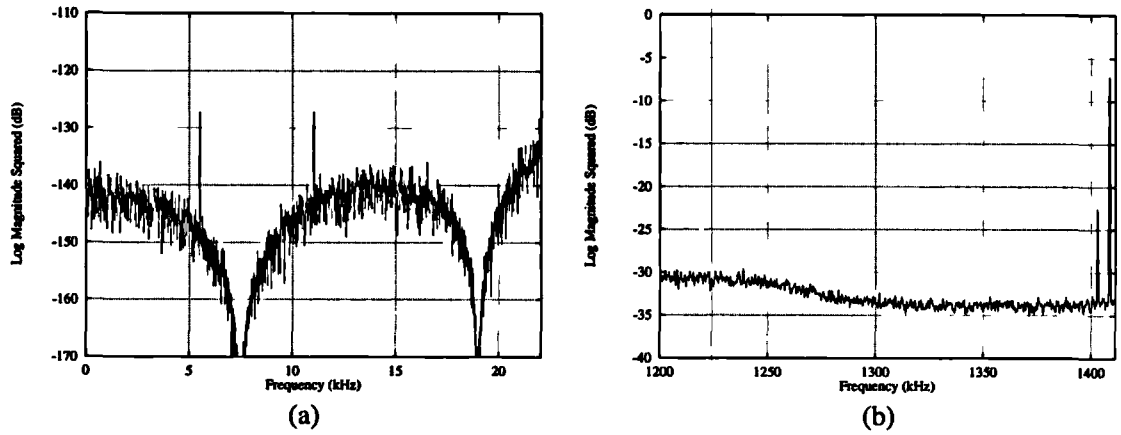


Figure 2.10: (a) Baseband and (b) Wideband spectra of example fourth order modulator with DC input $m_x = 1/512$

2.4.3 Noise Modulation

Noise modulation, is defined in [Van87] as the variation in noise power for slow changes in signal level. This may occur in practical implementations in the presence of DC drift and is therefore a greater problem with A-D converters than D-A converters.

The occurrence of noise modulation is linked to equation 2.9 and the Constant Output Power Property 2.2. Essentially this equation states that the output power is a constant which is divided between signal and noise power. For a slowly increasing DC input, the noise power must slowly decrease for equality to be maintained. The decrease in noise power usually occurs in conjunction with a change in the shape of the noise spectrum, due to a reduction in quasi-linear gain with increasing input amplitude. This characteristic variation in K_n will be discussed in greater detail in chapter 3. By equation 2.17 a reduction in K_n causes the *baseband* noise power to increase, despite the overall decrease in wideband noise power. The reduction in quantizer gain is more apparent at high input levels as the modulator approaches overload, therefore the noise modulation predicted by the quasi-linear model occurs predominantly at high levels.

High input-level noise modulation is not especially a problem in audio converters since quantization errors tend to be perceptually masked by the human ear when the input is a certain level above the quantization error. [Nor95].

A more serious phenomenon which is not predicted by the quasi-linear model is

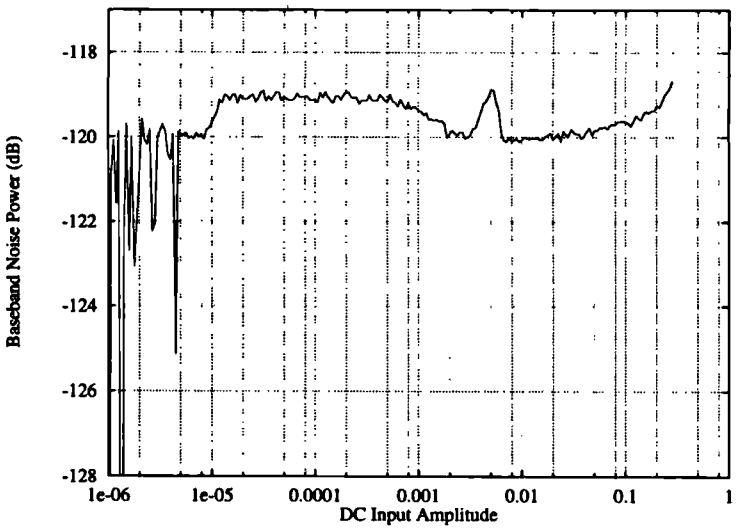


Figure 2.11: Noise Modulation plot of modulator {64, 4, 4.25}

the variation in noise power at low signal amplitudes. By equation 2.9, for a constant DC input amplitude, the error power is shared between noise and tone components, therefore an increase in out-of-band tonal power will generally result in a decrease in baseband noise power.

A typical noise modulation plot is shown in figure 2.11, for the modulator {64, 4, 4.25}. In this example, the modulator described in the previous example has been repeatedly simulated for DC inputs incremented in 0.5 dB steps. The simulation and parameter details are given in appendix A. In this example, severe noise modulation occurs for inputs below -100 dB, and there are regions in which the noise power falls to zero. These regions correspond to a spectrum which is purely tonal (i.e. there are no noise components), caused by the state space becoming locked onto short limit cycles with periods in the region of 50 samples. Example spectra will be presented in chapter 4 to demonstrate this.

2.4.4 Linearisation

Several schemes have been proposed to alleviate the idle tone problem and therefore enhance the linearity of the modulator. The most common are:

1. Adding a DC bias to the modulator input.

2. Dithering the quantizer.

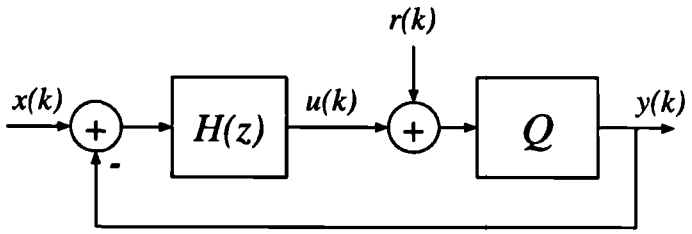


Figure 2.12: Block Diagram of Dithered ΣΔ modulator

3. Implementing chaotic modulation.

2.4.5 Adding a DC bias

The addition of a DC bias to the modulator input was one of the first proposed schemes for linearisation [Can76]. The bias is chosen to shift the idle tones away from the baseband. For an oversampling ratio L, from equation 2.23 the low frequency tone will only fall in the baseband if $m_x \leq 1/(2L)$. However, adding a DC bias does not remove the high frequency tone, which may intermodulate with any spurious clock frequencies to produce components in the baseband [Har92]. The scheme is also unsuitable where DC coding is required, unless the bias is accurately subtracted at the output [Can76]. Furthermore any DC component in the input signal will upset the bias setting, possibly causing tones to shift back into the baseband.

2.4.6 Dithering the Quantizer

Quantizer dithering is the traditional method of linearising systems employing quantization and has its roots in video coding [Van84]. The basic concept is to add a random or pseudo-random noise source to the quantizer input (figure 2.12). The initial work on dithering for audio systems was targetted at multibit quantizers. Vanderkooy and Lipshitz [Van87] have shown that the quantizer staircase function can be linearised by the introduction of rectangular probability density function (PDF) dither spanning one least significant bit (LSB) of the quantizer. This has the effect of removing nonlinear artifacts caused by the correlation between the quantizer error and its input. It is recognised [Van87] that rectangular PDF dither does not prevent noise modulation, but this can be eliminated by using triangular PDF dither spanning two quantizer LSBs [Van89].

Associated with the dithering of the quantizer is the introduction of wideband

noise. The performance of dither in this respect is measured in terms of the noise penalty, defined as the increase in baseband quantization noise power incurred by the inclusion of the dither source. A more general term is the dynamic range penalty, which is the reduction in dynamic range of the system with the inclusion of dither. The dynamic range penalties for optimally dithering a multibit quantizer with rectangular and triangular PDF dither are 3 dB and 4.8 dB, respectively [Van89]. Vanderkooy and Lipshitz also discuss triangular PDF high-pass dither, generated by the correlation of successive pairs of rectangular PDF dither samples [Van89]. The correlation is achieved by pre-filtering the rectangular PDF dither source with the transfer function $T(z) = 1 - z^{-1}$. For oversampled systems with multibit quantizers, the dynamic range penalty incurred by high-pass dither is smaller, since the dither power falling into the baseband is lower.

Dither can also be used for noise shaping coders [Van89] with dither generally applied within the feedback loop before the quantizer (figure 2.12). For a dither signal $R(z)$ and quantizer error $E_l(z)$, the output signal is given by:

$$Y(z) = X(z)\frac{H(z)}{1+H(z)} + E_l(z)\frac{1}{1+H(z)} + R(z)\frac{1}{1+H(z)} \tag{2.25}$$

This equation shows that the dither signal is noise shaped by the NTF. The dither may also be applied at the system input, if it is first passed through a filter implementing the NTF [Nor92].

The dithering of single-bit quantizers was first used in delta modulators [Van87]. The general goal of dithering single-bit quantizers is the elimination of limit cycles and the associated tones and noise modulation, rather than the linearisation of the quantizer transfer function.

A modulator dithered by an independent random noise source is no longer a deterministic system, therefore dither has the effect of breaking up limit cycles in the state-space. In [Ris94b], Risbo discusses the case of a periodic dither signal added to the quantizer input, and argues that the dither changes the composition of possible limit cycles. If the dither is random, a large number of temporary limit cycles can exist, though the random nature of the input prevents them from being sustained for any length of time. The increase in possible non-sustainable limit cycles tends to cause the output sequence to be randomized, resulting in a non-tonal spectrum.

For the dithering of A-D converters, it is difficult to implement a reliable analogue noise source using thermal noise because of temperature dependencies and the

requirements of a high gain amplifier which is well isolated from spurious clock signals [Ris94b]. An alternative method is to generate the dither signal digitally using a maximum length sequence and use a local D-A converter to convert the random signal to analogue. This method is inefficient to implement because an additional D-A converter is required.

2.4.7 Chaotic Modulation

An alternative method of linearisation which does not require a noise source is the implementation of modulators with unstable filter dynamics, i.e. chaotic modulation. This has received considerable interest, both from a practical and theoretical perspective. Chaotic modulation is achieved by moving one or more of the poles of $H(z)$ outside the loop filter so that $H(z)$ becomes open-loop unstable. As a consequence, small changes in the modulator input are amplified exponentially in time [Sch93]. In terms of the nonlinear dynamics of the modulator, chaotic modulators allow a large number of unstable limit cycles to exist [Ris94b]. Due to the unstable nature of the limit cycles, the output is generally non-periodic and the spectrum is continuous [Ris94b].

The moduli of poles outside the unit circle control the rate at which nearby points diverge and therefore the 'degree' of chaos [Ris94b]. If the poles are barely outside the unit circle, it may take a long time before a perturbed limit cycle diverges enough to cause a periodic pattern in the output to be disrupted. As a result, the output may appear periodic for a short time span, resulting in audible tones [Sch93].

Moving the poles of $H(z)$ outside the unit circle corresponds to moving the zeros of the NTF outside the unit circle and this can significantly increase the baseband noise power of the modulator. By the Noise Shaping Theorem, the NTFs of all chaotic modulators have an inequality in the areas above and below the 0 dB line, resulting in an increased NTF power gain and a poorer tradeoff between NTF attenuation and stability. For the *same* modulator stability, higher order chaotic modulators have considerably poorer noise performance than their dithered counterparts [Ris95].

An alternative method of implementing chaos with improved noise performance proposed by Risbo [Ris94b], involves the introduction of an all-pass term in the NTF. This comprises a real zero outside the unit circle and a reciprocal pole inside the unit circle, and has the advantage that the degree of chaos can be controlled

relatively independently of the noise spectrum, since the all-pass term ensures that the necessary scaling is unaffected. However the order of the loop filter is increased, resulting in a more complex implementation.

Due to the poor performance of chaotic modulation, in comparison to dithered modulation, chaos will not be considered further in this dissertation; however in chapter 4 an alternative linearisation scheme will be introduced, which is both efficient to implement and has relatively small dynamic range penalties, whilst being suitable for A-D converter implementations.

2.5 Architectures for Adaptive Modulation

In this section we present a brief overview of techniques and architectures which fundamentally modify the operation of modulator in order to gain dynamic range advantages. This section forms the background to chapter 5. As explained in section 2.3.1, a fundamental trade-off exists between the stability of the modulator and the noise in the baseband. The techniques described here improve this trade-off by making the modulator adaptive in some way. An adaptive system is one which tracks in a useful manner some feature of its external environment. In the case of adaptive modulators, generally either the quantizer or loop filter is adapted according to the system input level. Adaptive systems are classified as either forward or backward adaptive [Yu92] according to whether information from the system input or output is used to control the adaption. For $\Sigma\Delta$ modulators both input and output signals contain information about the input level, therefore the choice of forward or backward adaption is usually related to ease of implementation. The adaption may also be either instantaneous or block-wise, where the adaption occurs on every sample or after every block of samples, respectively [Yu92], and this choice influences the speed at which the adaption can respond to the environment.

2.5.1 Adaptive Loop Filter

Adaptive Coefficients

A $\Sigma\Delta$ modulator using an adaptive loop filter has been described by Yu [Yu92], in which the NTF poles are adapted according to the input level. At high input levels, the poles are moved close to the baseband. This reduces the transition band, so

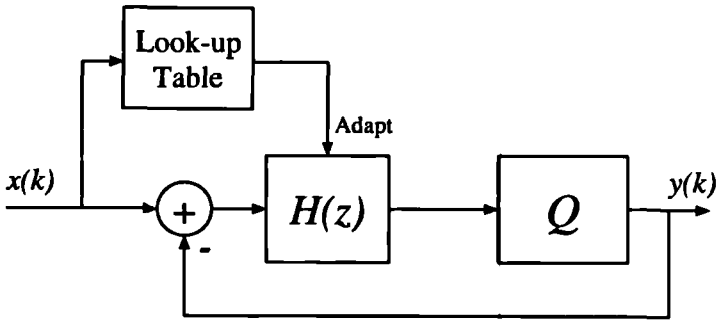


Figure 2.13: Adaptive Filter due to Yu, shown with Instantaneous Forward Adaption

that after necessary scaling the power gain of the filter is low and by equation 2.15, a higher input level can be accepted. At low input levels, the poles are moved away from the baseband, causing the power gain to increase but the baseband noise power to reduce. In this way dynamic range is enhanced, at the expense of input-dependent baseband noise power.

The adaption scheme involves dividing the dynamic range of the input signal into six regions, with each region having a corresponding NTF. In each region the NTF is designed to have the lowest possible baseband noise power yet still remain stable for all inputs within the region. An optimisation procedure based upon a direct search algorithm is used to find the coefficients. Schemes based upon block and instantaneous adaption have been investigated and in both cases a look-up table is used to map the input range onto coefficient sets. The block diagram of the adaptive modulator using instantaneous forward adaption is shown in figure 2.13.

A disadvantage with this scheme is that unless the baseband attenuation is conservatively selected, unstable regions may exist within the amplitude regions, which are not detected during optimisation [Dun96]. A scheme which detects instability and automatically adapts the filter will be described in chapter 5.

Clipping

An alternative form of adaptive filter uses nonlinear elements which limit the magnitude of the modulator state-space variables upon detection of an overload condition, by either reset or clipping. Such a scheme prevents the occurrence of self-sustaining saturation limit cycles which occur in higher order modulators. A similar scheme has been used in delta modulators [Ada84]. Clipping or reset causes state space information to be lost. As a consequence, the noise shaping mechanism which causes

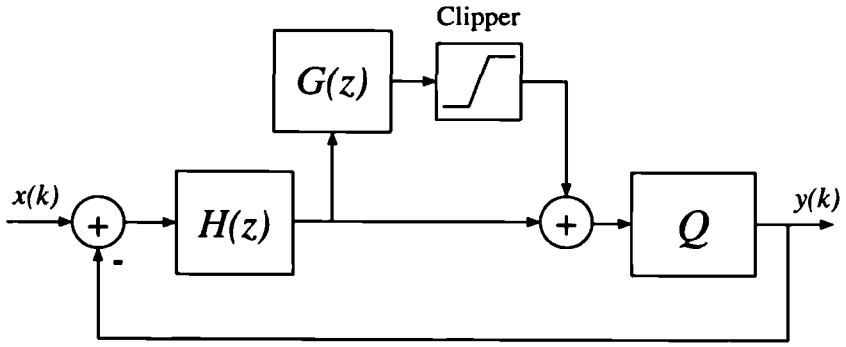


Figure 2.14: Modulator with Selective Clipping

error cancellation in the baseband is disturbed and the baseband noise power increases considerably whilst the input level is maintained. A method of limiting such loss of information described by Naus [Nau91] and also by Thurston [Thu94] uses a modified limiting scheme in which higher order sections are removed from the loop upon the detection of an overload condition. In figure 2.14 a general block diagram of the technique is shown. When clipping occurs, the loop reverts to the second order $H(z)$ and stability is maintained without a complete loss of state information. The technique described in chapter 5 is conceptually similar to this.

2.5.2 Adaptive Quantizer

Adaptive quantizer step size

In adaptive quantization, the amplitude of either the quantizer step or its feedback component is adapted in order to match the variance of its input signal to avoid overload. The method was first used in Delta Modulation to avoid slope overload distortion without incurring excess granular noise [Ste75] and was investigated extensively for $\Sigma\Delta$ modulators by Yu [Yu92] (figure 2.15). Although dynamic range advantages are gained from adapting the step size, there are implementation problems. If the adaption is in the feedback path, a decoding network is required after the modulator. If the adaption is in the feedforward path, a multilevel quantizer is required to implement the different step sizes, and the advantages of one-bit conversion are lost.

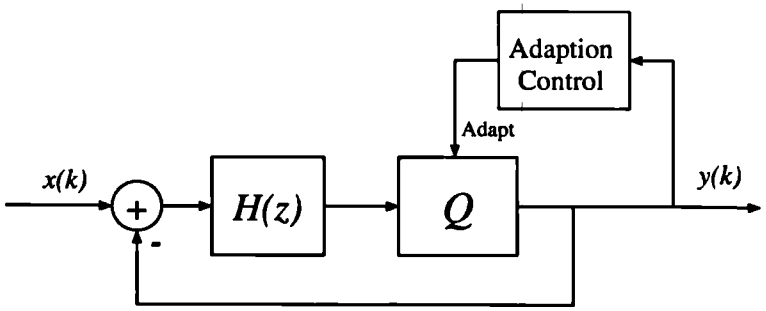


Figure 2.15: Adaptive Quantizer due to Yu, shown with adaption in the feedforward path

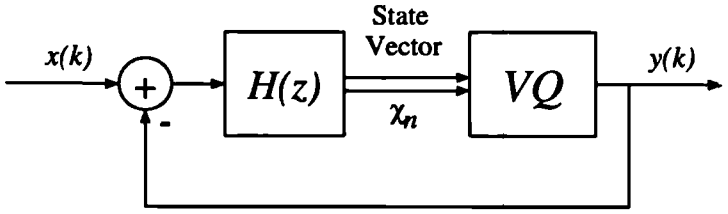


Figure 2.16: Modulator with Vector Quantization, due to Risbo

Vector Quantization

A scheme which retains single-bit quantization and requires no special decoding uses a modified quantizer termed a Vector Quantizer [Ris93]. A general vector quantizer maps the filter states nonlinearly onto the quantizer, so that more information from the state space can be used in the quantization process [Ris94b] (figure 2.16). The Vector Quantizer is used to improve stability by preventing the modulator from entering regions in state space corresponding to saturation limit cycles. To achieve this the algorithm must detect instability, then return to a point in state space before the instability occurred. The unstable region is then avoided by inverting the quantizer state so that the modulator enters a different and (potentially) stable region in state-space. The return to a stable point in state-space involves restarting the modulator from a previous time-step, which is difficult to implement in practice. The concept of modifying the quantizer state is very useful, however, for the reasons described in section 2.7.

2.6 Power Digital-to-Analogue Conversion

This section forms the background to chapter 6. Power digital-to-analogue conversion is an application of single-bit converters which has received relatively little attention. The basic concept is to replace the low-level sample-and-hold and low-pass filter by a high-level power switch and filter (figure 1.5). The power switch typically consists of a MOSFET bridge circuit, producing a rectangular waveform typically in the range $\pm 30 - 50$ volts. The low-pass filter is a passive LC network. The power-switch and filter combination is commonly known as a class-D amplifier. The combination of the one-bit converter and class-D amplifier is referred to as a power D-A converter, or sometimes as a digital power amplifier. The two principal advantages of power D-A converter for audio applications, in comparison to the classical low-level D-A converter followed by analogue amplification, are:

1. Removal of the analogue amplifier and elimination of associated distortions, for example class B crossover distortion.

2. High power efficiencies. Class-D amplifiers have a theoretical efficiency of 100%. Practical designs can achieve efficiencies upwards of 90% [Ped94]. This can lead to a compact amplifier, which requires very little heatsinking [Hi-95].

Of course, power D-A converters may also be used in other applications where conversion to analogue at high power is required, for example motor and plant control, and also portable systems, where high efficiencies are of paramount importance, for example personal communication systems and hearing aids. The work on power D-A converters presented in this dissertation concentrates on the one-bit converter, rather than the class-D amplifier. Non-idealities in the class-D amplifier do influence the system performance, however, and sensitivity to these can in some cases be reduced by appropriate converter design. Two non-idealities which this applies to are:

1. Finite rise and fall times. These influence the performance of the system in two respects. Firstly, energy dissipation occurs whenever current flows through a device which has a voltage drop across it. This will occur on every transition, and so to minimise power dissipation, the transition rate must be kept to a minimum. The transition rate is usually described in terms of the average pulse-repetition frequency (PRF), which is defined as the reciprocal of the

average time between consecutive rising edges of the one-bit signal. Second, any mismatch between rise and fall time can cause a signal-dependent error, causing nonlinear distortion. It will be shown in appendix E.2 that this can be eliminated by ensuring the PRF is constant.

2. Timing jitter. This is a system clock timing error in which the practical clock instants deviate from their ideal times, generally as a result of noise in the clock generator circuit [Pau95]. Since the timing of the clock only influences the output waveform on the occurrence of transitions, it is possible to minimise the influence of jitter by keeping the PRF as low as possible.

The ideal one-bit converter will therefore have a low and constant PRF. Reports (for example [Ped94]) suggest a PRF around 350 kHz is acceptable for efficient power conversion.

2.6.1 Pulse-Width Modulation Power D-A Converters

The majority of research into power D-A converters have concentrated on the use of converters with pulse-width modulation output stages. The use of PWM power D-A converters for audio was proposed by Sandler in 1983 [San83] using initially a simple scheme in which each PCM input word is mapped onto an equivalent pulse-width. This is termed uniformly sampled PWM. Uniformly sampled PWM has three distinct disadvantages, however, which have been recognised throughout the literature. Firstly, it is inherently a nonlinear process, generating harmonic distortion and a unique distortion termed foldback distortion, where sidebands of the carrier frequency (i.e PRF) fall into the baseband. These distortions are detailed in the paper of appendix E.1. Secondly, due to the one-to-one mapping between pulse-width and pulse amplitude, the time-domain pulse resolution required to achieve 16-bits amplitude resolution is 2^{16} greater than PCM, and so a 44.1 kHz sample rate signal requires a resolution corresponding to a clock frequency of $44100.2^{16} = 2.89\ GHz$. This places extreme and unrealisable demands on the speed of the switching stage. Thirdly, guard bands where no pulses are allowed are required to ensure that adjacent pulses are well separated, in order to allow the output stage to recover between transitions [Hio93b]. This complicates the design of the PWM stage.

These problems have led to the use of architectures using oversampling to reduce foldback distortion [San83], noise shaping to reduce the PCM wordlength and the associated clock frequency [Gol89], and a variety of schemes to reduce harmonic distortion [Lei90, Gol94, Haw92, Cra93, Ped94]. These schemes are detailed in the paper of appendix E.1. Even with the use of these schemes, very careful noise transfer function design is required to reduce *foldback noise*, where out-of-band noise components intermodulate in the nonlinear PWM stage and corrupt the baseband [Pau95].

Typical PWM systems use eight times oversampling and noise-shaping to reduce the wordlength at the input of the PWM block to eight bits, corresponding to a PRF of 352.8 kHz. A clock frequency of approximately 90 MHz is required and this is beyond the frequency of standard ASIC implementation, therefore current implementations use a discrete logic circuit to generate the PWM output [Hio94]. From an implementation perspective, linearised noise shaped PWM is not ideal; however this has not prevented the development of a commercial prototype [Hi-95] which is based upon a published design capable of delivering 600W into a four ohm load with an efficiency more than 90% at maximum power level [Ped94].

2.6.2 Sigma-Delta Power D-A Converters

Alternative $\Sigma\Delta$ modulation based power D-A converters have been largely discounted due to the high rate of transitions in the bitstream, which can theoretically reach half the clock frequency ($\approx$ 1.4 MHz for a 64 times oversampled system). $\Sigma\Delta$ modulation does have two distinct advantages, however. Firstly it is capable of highly linear conversion (assuming the problems of idle tones are overcome). Secondly the clock frequency is considerably lower that that of PWM, allowing the highly desirable possibility of ASIC implementation and eliminating the need for guard bands. The problems associated with the high transition rate have been addressed in [And93], by using an alternative output stage which operates by resonant mode where current and voltage pulses are non-simultaneous. The performance of this system is inferior to Class-D amplification in terms of noise, distortion and efficiency [Ped94]. In chapter 6, new techniques are used to reduce the transitions in the bitstream, allowing the use of conventional Class-D amplification and a modulator suitable for ASIC implementation.

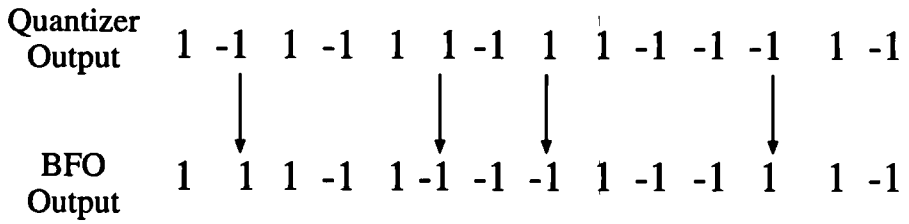


Figure 2.17: Concept of Bit-flipping

2.7 Bit-Flipping Modulators

The fundamental contribution of this thesis is the introduction of modulator architectures and algorithms based upon a new technique termed bit-flipping. The concept of bit-flipping hinges on a modified quantizer, whose output state is selectively inverted so that the bit sequence at the output of the modulator is modified (figure 2.17). The selective inversion is performed by a *bit-flipping operator* (BFO) which has the following transfer function:

$$y(k) = \begin{cases} -b(k) & \text{With bit-flipping} \\ b(k) & \text{Without Bit-flipping} \end{cases} \tag{2.26}$$

The general block diagram of the bit-flipping modulator is shown in figure 2.18. The BFO can be conceptually thought of as a multiplier which selectively inverts the quantizer output by multiplying it by plus or minus one on every sample, according to the value of its control input. The BFO is located within the feedback loop, so that the modification of the quantizer decision is corrected in subsequent samples. This significantly influences the stability of the modulator. The control input of the BFO is connected to the output of the control algorithm, which generates either plus or minus one on every sample, according to a nonlinear combination of its inputs, which may optionally include the quantizer input and an external input $c(k)$ (for example a random condition). Since the decision made by the quantizer is modified for some samples according to additional information, bit-flipping is related to Vector Quantization [Ris94b].

Bit-flipping may be controlled in order to modify the bit-patterns which occur at the modulator output. For example, it may be used to break up limit cycles which occur at the modulator output in order to enhance the linearity of the conversion (refer to chapter 4). By modifying the quantizer decision in a manner which is in some way dependent on the input signal level, bit-flipping modulators can also be

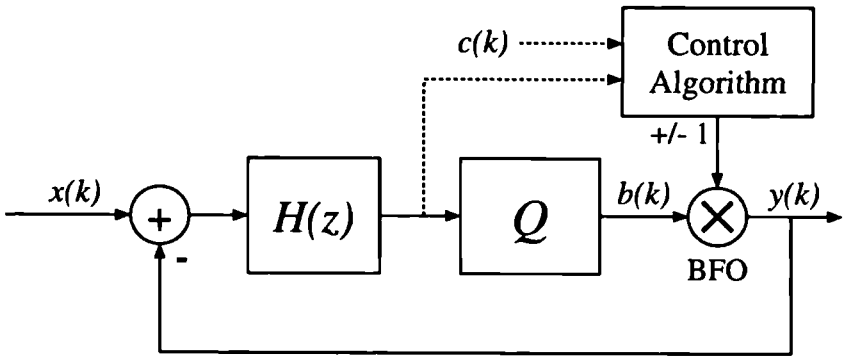


Figure 2.18: Architecture of Bit-Flipping Modulator

made adaptive (refer to chapter 5). A significant advantage of bit-flipping modulators over the other adaptive quantizer schemes is that the quantizer output remains single bit therefore no special decoding network is required and the linearity of one-bit conversion is retained. Furthermore, bit-flipping is efficient to implement as it represents only a sign inversion.

Bit flipping may also be used to force grouped patterns to occur, and this has applications in power D-A converters (refer to chapter 6).

2.8 Scope of Investigation

The majority of the work presented investigates new architectures and algorithms based upon the concept of bit-flipping. It has been the intention to investigate the applicability of the technique to a wide range of modulators parameters i.e. order, oversampling ratio and power gain.

For practicality it has been necessary to restrict the scope of investigation. The main application is for audio converters, and in this field practitioners have largely discounted first and second order modulators due to the correlated quantization noise and tonal artifacts [Ada91]. Furthermore, low order modulators require a high oversampling ratio for good noise performance, which can lead to undesirably high clock frequencies. Accordingly, the investigation concentrates on third and higher order modulators.

To further restrict the required range of simulations, most results use an oversampling ratio of 64 times, and universally the sampling frequency is 44.1 kHz, which is used for compact disc recordings. The use of 64 times oversampling is

typical of audio quality converters (see for example [Ada91, Well89, Iso91]). Where appropriate, other oversampling ratios are investigated, for instance in the case of power D-A converters (chapter 6) where the bitstream rate is crucial to the power dissipation in the output stage.

The simulation scheme used in evaluating modulator performance is described in appendix A.1. The aim has been to establish the fundamental character of the modulation process, therefore wordlength and coefficient quantization effects have not been considered.

For all investigations the NTFs have been designed using the filter class described in section 2.3.1. A set of NTFs have been obtained with cutoff frequencies yielding power gains in the range 1 $dB \rightarrow 5$ dB after necessary scaling. This range has been found empirically to cover modulators ranging from highly stable (i.e. an ability to accept an input approaching full scale) to completely unstable (even at zero input). The notation used to describe the modulator parameters is defined in appendix A.2.

2.9 Summary

In this chapter the relevant background work in five main areas of $\Sigma\Delta$ modulation has been summarised: modelling, design, linearity, adaption and power conversion.

The Discrete-time model of the $\Sigma\Delta$ modulator has been introduced. The noise transfer function (NTF) and signal transfer function (STF) have been defined and it has been shown how these relate the loop filter to the noise and signal components in the modulator output. The characteristics of the quantizer error signal have been discussed in relation to the assumption that it can be modelled as an independent white noise source.

The quasi-linear model has been introduced, which models the quantizer as a gain term followed by an additive error. The quantizer gain and error source are defined according to loop statistics and are therefore input signal dependent. This allows observed changes in the quantization noise with input level to be modelled. Simple expressions have been given, which relate the baseband noise power to the quantizer gain and NTF.

The issue of modulator stability has been discussed and the problems of conventional stability definitions highlighted. A definition of stability has been proposed which states that a modulator is stable if it accurately codes a DC input. It has

been shown that stability relates to the power gain of NTF and that DC coding may fail if the input level is too high.

Some of the issues of designing NTFs have been reviewed. An important theorem is the Noise Shaping Theorem, which provides a link between the transition width and the power gain of the NTF. Therefore a tradeoff exists between the baseband noise power and maximum stable amplitude (MSA). A method of designing NTFs for baseband coding has been described, using the high-pass Butterworth class of filters. These design techniques are used for all the modulators in subsequent chapters.

Certain nonlinear aspects of $\Sigma\Delta$ modulator conversion have been discussed with reference to the performance in an audio context. The main problem is the presence of idle tones, which are linked to limit cycles in the state-space of the system. Expressions have been derived for dominant low and high frequency tones which occur for a rational DC input in a general modulator. Both tones are problematic, since the low frequency tone may fall in the baseband, and the high frequency tone can intermodulate back to the baseband in a real implementation. It has also been shown that idle tones can cause noise modulation, where the baseband noise power varies with input level. In audio converters, noise modulation can be psychoacoustically disturbing.

Three of the common linearisation techniques have been reviewed. The first method is the addition of a DC bias to the modulator input. This causes the low frequency tone to shift out of the baseband, but the technique is unsuitable for DC coding and does not eliminate the high frequency tone. The second method is the addition of a noise source to the quantizer input, termed dither. For $\Sigma\Delta$ modulators, the dither signal changes the composition of possible limit cycles, and tends to randomise the output noise spectrum . A disadvantage of dither is that it is difficult to implement in A-D converters. The third technique, involves modifying the loop filter so its poles are outside unit circle and the modulator becomes chaotic. The main advantage of chaos is that no analogue noise source is required in A-D converter implementations. A significant disadvantage, however, is that higher order modulators linearised with chaos have significantly poorer dynamic range than dithered modulators.

Some of architectures for adaptive modulation have been discussed, including adaptive loop filters, adaptive quantization and vector quantization. The technique of adapting the loop filter is particularly interesting from a theoretical viewpoint - it

causes the transition width of the NTF to reduce at high input levels so that power gain is reduced and the stability of the modulator is enhanced.

The concept of using $\Sigma\Delta$ modulation for power D-A conversion has been introduced. Power D-A converters have high efficiency and potentially low complexity and distortion. Previous work in this field has concentrated on pulse-width modulation (PWM) techniques. The two main problems with PWM are its poor linearity and high clock frequencies required to time the pulse edges. As a consequence the system complexity is relatively high and ASIC implementation is not possible. The $\Sigma\Delta$ modulator can potentially improve in these areas because of its high linearity (when used with an appropriate linearisation scheme) and considerably lower clock frequency.

Finally, the concept of the bit-flipping modulator has been introduced, Bit-flipping is the selective inversion of the quantizer output state. It is implemented with a bit-flipping operator which is located within the feedback loop to ensure that the coding accuracy is not destroyed. Following a detailed study of quasi-linear analysis, it is the purpose of chapters 4, 5 and 6 to show how bit-flipping may be applied to linearisation, adaption and power conversion.

Chapter 3

Quasi-Linear Modelling

3.1 Introduction

In this chapter the technique of quasi-linear modelling is used to analyse $\Sigma\Delta$ modulators. The aim is to present an analysis framework which gives an insight into the noise and stability behaviour of $\Sigma\Delta$ modulators and to describe how the behaviour relates to the input signal and NTF.

The quasi-linear method is not formulated towards precise analysis due to the use of assumptions about certain signal characteristics. The main advantage lies in its general applicability to both low and high order modulators (defined in section 2.2.3) and modified architectures such as dithered or bit-flipping modulators.

The concept of quasi-linear analysis has been introduced in section 2.2.2. To summarise, the quantizer is approximated by a linear gain term followed by an additive white noise source. The gain and noise variance are signal-dependent parameters, defined in the steady-state according to the statistics of the quantizer input and output signals. A formal method of defining the parameters has been described by Ardalan and Paulos [Ard87]. The method originates from the describing function technique of Booton [Boo53], used in the analysis of control systems.

A key feature of the modelling is that it allows different nonlinearities to be dealt with in a straightforward manner and this will be of considerable use in chapter 4 for the modelling of dithered and bit-flipping modulators.

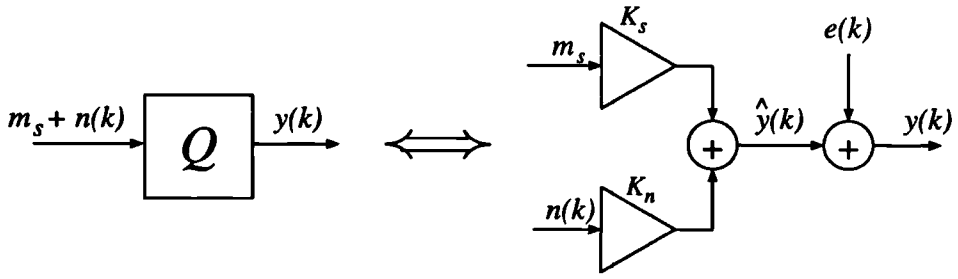


Figure 3.1: Modelling the quantizer with separate quasi-linear gains for the signal (K_s) and noise (K_n). The parameters $n(k)$ and m_s are related to $u(k)$ by equation 3.1

3.2 Formulation of the Quasi-Linear Model

In this section we consider quasi-linear modelling in detail for a $\Sigma\Delta$ modulator with a DC input m_x. The primary reason for using a DC input is that the analysis is greatly simplified, although useful results can still be obtained which are relevant to a slowly varying input signal (refer to section 2.2.3).

The quasi-linear model deals with DC inputs by separating the quantizer input $u(k)$ into AC and DC components, $n(k)$ and m_s

$$u(k) = n(k) + m_s \tag{3.1}$$

The two components are passed through separate noise and signal gain terms (K_n and K_s respectively), which are combined to form the signal $\hat{y}(k)$:

$$\hat{y}(k) = n(k)K_n + m_s K_s \tag{3.2}$$

The use of separate gain terms makes the analysis easier, and the signal gain K_s is eliminated from the final expressions.

Ardalan and Paulos define $\hat{y}(k)$ as an *estimate* of the quantizer output, which differs from the real quantizer output by an error $e(k)$, as shown in figure 3.1.

Therefore the error $e(k)$ can be considered as the modelling error, defined as the difference between the output of the actual modulator $y(k)$, and the estimate $\hat{y}(k)$.

$$\begin{aligned} e(k) &= y(k) - \hat{y}(k) && (3.3) \\ &= y(k) - n(k)K_n - m_s K_s && (3.4) \end{aligned}$$

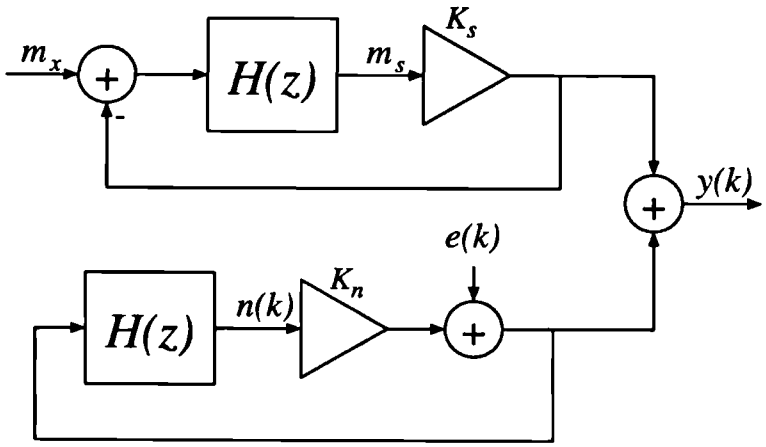


Figure 3.2: Quasi-Linear Model

An essential part of the modelling is to choose the two gains so as to minimise the variance of the modelling error $\sigma_e^2 = E\{e^2(k)\}$. This choice can also be shown to decorrelate $e(k)$ from $n(k)$ [Ris94b]. This means that it is now possible to include the error in the model as an independent noise source. Of course in practice, this error is the noise introduced by the quantizer.

The system can be redrawn as two interlocking systems, with separate loops handling the AC and DC components [Ard87], as shown in the complete model of figure 3.2. The output is now given as:

$$y(k) = n(k)K_n + m_s K_s + e(k) \tag{3.5}$$

The circulating AC component in the loop is generated by the error $e(k)$. This is modelled as an external noise source in figure 3.2 but in the real modulator it is generated internally by the quantizer. The DC component is generated by the system input m_x.

3.2.1 Alternative Definition of Quantizer Gain

It is worth noting that different definitions of the quantizer gain K_n also appear in the literature. In [Hei93a], a quantizer describing function method is used, in which the transfer function to sinusoidal inputs is approximated, which derives from work by Booton [Boo53]. In [Ada91], the statistical describing function technique is chosen, but the ratio of the mean of the quantizer input and output is used to define K_n. Again this technique is used in control theory [Tha62]. In [Sti88b] the

gain K_n is chosen by equating the quantizer output power with a known input power an an assumed quantizer error variance. This is used to describe the stability of the modulator by means of the root-locus technique. An identical method of defining K_n is used in [Magu94]. A different approach is used in [Rit91] and [Bai93], in which K_n is defined on a sample-by-sample basis as the ratio of the quantizer output and input. This also allow the stability of the modulator to be characterised using root-locus analysis, but is not formulated towards noise analysis.

3.3 Evaluation of Quasi-linear Parameters

Using the definitions of Ardalan and Paulos, it is possible to derive the quasi-linear parameters K_n and σ_e^2 either in terms of the probability density function (PDF) of the AC component of the quantizer input, or in terms of time averages. The former method requires an assumption to be made about the PDF of the quantizer input. The latter makes no such assumption, but requires the parameters to be determined by direct simulation of the modulator.

3.3.1 Probability Density Function Method

First the PDF method will be described and an assumption will be made about the quantizer input PDF observed in a real modulator. In [Ard87] it is assumed that the observed PDF has a Gaussian distribution. This is reasonable because the quantizer input signal is taken from the filter output which, due to the additions in the filtering, will tend towards a Gaussian distribution by the Central Limit Theorem [Ard87]. This assumption is fundamental to the accuracy of the modelling and any errors will influence the accuracy of the final solution.

First, we find expressions for the quasi-linear parameters. From equation 3.1 the quantizer output can be expressed as:

$$y(k) = \mathcal{Q}(n(k) + m_s) \tag{3.6}$$

where $\mathcal{Q}()$ is the quantizer nonlinearity i.e. the signum function

$$\mathcal{Q}(n(k) + m_s) = \begin{cases} \Delta & n(k) > -m_s \\ -\Delta & n(k) \leq -m_s \end{cases} \tag{3.7}$$

and $\pm\Delta$ are the quantizer output levels (normally defined as ± 1).

Equation 3.4 can therefore be written as:

$$e(k) = \mathcal{Q}(n(k) + m_s) - n(k)K_n - m_sK_s \tag{3.8}$$

In the following integral equations, the bracketed k terms are dropped to simplify notation. Assuming ergodicity, time averages can be described by statistical moments, therefore the error variance σ_e^2 is found by

$$\begin{aligned}\sigma_e^2 &= \int_{-\infty}^{\infty} e^2 p(n)dn \\ &= \int_{-\infty}^{\infty} \{\mathcal{Q}(n + m_s) - nK_n - m_sK_s\}^2 p(n)dn\end{aligned}$$

K_n and K_s are chosen to minimise the variance of σ_e^2, by partial differentiation.

$$\frac{\partial}{\partial K_n}\{\sigma_e^2\} = 0 \tag{3.9}$$

$$\frac{\partial}{\partial K_s}\{\sigma_e^2\} = 0 \tag{3.10}$$

K_n is evaluated as:

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial K_n} \{\mathcal{Q}(n + m_s) - nK_n - m_sK_s\}^2 p(n)dn = 0$$

$$K_n \int_{-\infty}^{\infty} n^2 p(n)dn - \int_{-\infty}^{\infty} \mathcal{Q}(n + m_s)np(n)dn + m_sK_s \int_{-\infty}^{\infty} np(n)dn = 0$$

Since $E\{n(k)\} = 0$

$$K_n = \frac{1}{\sigma_n^2} \int_{-\infty}^{\infty} \mathcal{Q}(n + m_s)np(n)dn \tag{3.11}$$

where $\sigma_n^2 = \int_{-\infty}^{\infty} n^2 p(n)dn$
Similarly, K_s is found.

$$K_s = \frac{1}{m_s} \int_{-\infty}^{\infty} \mathcal{Q}(n + m_s)p(n)dn \tag{3.12}$$

Since $\int_{-\infty}^{\infty} \mathcal{Q}(n + m_s)p(n)dn$ is equal to the mean modulator output m_y:

$$K_s = \frac{m_y}{m_s} \tag{3.13}$$

Now the fundamental assumption is made that n has a Gaussian distribution.

$$p(n) = \frac{1}{\sigma_n\sqrt{2\pi}} e^{-n^2/2\sigma_n^2} \tag{3.14}$$

At this stage the nonlinearity is evaluated from equation 3.7.

$$\begin{aligned} K_n &= \frac{\Delta}{\sigma_n^3\sqrt{2\pi}}\left\{\int_{-m_s}^{\infty} ne^{-n^2/2\sigma_n^2}dn - \int_{-\infty}^{-m_s} ne^{-n^2/2\sigma_n^2}dn\right\} \\ &= \frac{2\Delta}{\sigma_n\sqrt{2\pi}}e^{-m_s^2/2\sigma_n^2} \end{aligned} \tag{3.15}$$

and for the signal gain:

$$\begin{aligned} K_s &= \frac{\Delta}{\sigma_n m_s\sqrt{2\pi}}\left\{\int_{-m_s}^{\infty} e^{-n^2/2\sigma_n^2}dn - \int_{-\infty}^{-m_s} e^{-n^2/2\sigma_n^2}dn\right\} \\ &= \frac{\Delta}{m_s}erf\left(\frac{m_s}{\sigma_n\sqrt{2}}\right) \end{aligned} \tag{3.16}$$

where $erf(y)$ is the error function:

$$erf(y) = \frac{2}{\sqrt{\pi}}\int_0^y e^{-u^2}du \tag{3.17}$$

Error Variance

The Constant Output Power Property (2.2) is now used to determine the noise power injected by the quantizer i.e. the variance σ_e^2. From equation 3.5 the quantizer output variance is given by:

$$\begin{aligned} E\{y(k)^2\} = \Delta^2 &= E\{e(k)^2\} + K_n^2E\{n(k)^2\} + m_y^2 + 2K_nE\{e(k)n(k)\} \\ &+ 2m_yE\{e(k)\} + 2K_nm_yE\{n(k)\} \end{aligned} \tag{3.18}$$

It can be shown that $e(k)$ is uncorrelated with $n(k)$ i.e $E\{e(k)n(k)\} = 0$. Furthermore $E\{e(k)\} = E\{n(k)\} = 0$ and so:

$$\sigma_e^2 = \Delta^2 - K_n^2\sigma_n^2 - m_y^2 \tag{3.19}$$

Substituting K_n from equation 3.15:

$$\sigma_e^2 = \Delta^2\left(1 - \frac{2}{\pi}e^{-m_s^2/\sigma_n^2} - \frac{m_y^2}{\Delta^2}\right) \tag{3.20}$$

From equations 3.13 and 3.16 a relationship is found between the quantizer input mean m_s and variance σ_n^2

$$\frac{m_s^2}{\sigma_n^2} = 2\left[erf^{-1}\left(\frac{m_y}{\Delta}\right)\right]^2 \tag{3.21}$$

therefore

$$\sigma_e^2 = \Delta^2\left(1 - \frac{2}{\pi}e^{-2\left[erf^{-1}(m_y/\Delta)\right]^2} - \frac{m_y^2}{\Delta^2}\right) \tag{3.22}$$

This expression reveals the property that, for a Gaussian distributed quantizer input, the error variance is independent of the NTF and depends only on the ratio of the mean modulator output and quantizer level. Note that this expression relies only on the assumption that the AC component of the quantizer input has a Gaussian distribution. No assumptions have been made about the spectral properties of σ_e^2.

Quasi-linear Gain

The evaluation of K_n is now considered. From equations 3.15 and 3.21

$$K_n = \frac{2\Delta}{\sigma_n\sqrt{2\pi}}e^{-2\left[erf^{-1}(m_y/\Delta)\right]^2} \tag{3.23}$$

From figure 3.2, the z-transform of the signal $n(k)$ is:

$$N(z) = E(z)\frac{H(z)}{1 + H(z)K_n} \tag{3.24}$$

The assumption is now made that $e(k)$ is a white noise source, therefore σ_n^2 is linked to σ_e^2 by the power gain of $H(z)/(1 + H(z)K_n)$.

Defining:

$$S_k(z) = \frac{H(z)}{1 + H(z)K_n} \tag{3.25}$$

The power gain is a function of K_n, given by

$$P_s(K_n) = \frac{1}{\pi}\int_0^\pi |S_k(\theta)|^2 d\theta \tag{3.26}$$

$$= \sum_{k=0}^{\infty} s_k(k)^2 \quad \text{by Parseval's relation} \tag{3.27}$$

Therefore the relationship is obtained:

$$\sigma_n^2 = \sigma_e^2 P_s(K_n) \tag{3.28}$$

Substituting equation 3.28 into equation 3.23:

$$K_n = \frac{2\Delta}{\sigma_e^2 P_s(K_n)\sqrt{2\pi}}e^{-2\left[erf^{-1}(m_y/\Delta)\right]^2} \tag{3.29}$$

which, for given σ_e^2, $H(z)$ and m_y, may be expressed as :

$$f(K_n) = 0 \tag{3.30}$$

The solution of this equation may be obtained using a nonlinear equation solver, and this approach is described in appendix B.2 and used to obtain results in section 3.4. The existence of a solution for different values of the mean output signal m_y will be discussed in section 3.5.

Evaluation of Baseband Noise Power

The solutions of σ_e^2 and K_n can then be used to evaluate the baseband noise power of the converter, using equation 2.5, in chapter 2:

$$P_b = \frac{\sigma_e^2}{\pi} \int_0^{\pi/L} \left| \frac{1}{1 + H(\theta)K_n} \right|^2 d\theta \tag{3.31}$$

where the assumption is made that $e(k)$ is a white noise source. This can be evaluated numerically using Simpson's Rule.

Results will be presented in section 3.4 and 3.6 to compare the baseband noise power obtained using this model and the value obtained in simulation.

3.3.2 Time-Average Method

In this section an alternative method of evaluating the quasi-linear parameters K_n and σ_e^2 is discussed, which requires no assumption to be made about the PDF of the AC component of the quantizer input. In this method, termed the time-average method, the quasi-linear parameters are evaluated during a discrete-time simulation of the modulator. The appropriate time averages are now derived, beginning again from equation 3.4:

$$e(k) = y(k) - n(k)K_n - m_s K_s \tag{3.32}$$

The variance σ_e^2 is expressed as

$$\sigma_e^2 = E\{[y(k) - n(k)K_n - m_s K_s]^2\} \tag{3.33}$$

and K_n can be determined as the value which minimise σ_e^2

$$\frac{\partial \sigma_e^2}{\partial K_n} = -2E\{n(k)y(k)\} + 2K_n E\{n^2(k)\} = 0 \tag{3.34}$$

therefore

$$K_n = \frac{E\{y(k)n(k)\}}{E\{n^2(k)\}} \tag{3.35}$$

From equation 3.1:

$$K_n = \frac{E\{u(k)y(k)\} - m_s m_y}{E\{u^2(k)\} - m_s^2} \tag{3.36}$$

σ_e^2 is evaluated from equation 3.19 i.e.

$$\sigma_e^2 = \Delta^2 - K_n^2 \left[E\{u^2(k)\} - m_s^2\right] - m_y^2 \tag{3.37}$$

The values of K_n and σ_e^2 may be evaluated from the time averages in equations 3.36 and 3.37 using a computer simulation, allowing the baseband noise power of the modulator to be calculated from equation 3.31.

3.4 Examples of Quasi-linear Analysis

In this section we present examples for the evaluation of the quasi-linear parameters and the resulting baseband noise prediction using the PDF and time-average methods for an example modulator with parameters {64, 4, 3.5} (refer to appendix A.2). For the simulations, 100,000 samples are used, with an initial offset of 1000 samples. These parameters have been found to yield consistent results i.e. increasing the offset or number of samples does not significantly change the time-averages. Further results will be given in section 3.6.

First of all we consider the variance σ_e^2 which represents the error introduced by the quantizer. Its value is given by equation 3.22 and depends only upon the ratio of the mean quantizer output m_y and the quantizer level Δ. The error variance $\sigma_e^2(m_y)$ with $\Delta = 1$ is plotted in figure 3.3 for the two methods. For the PDF method, $\sigma_e^2(m_y)$ reduces to zero for $m_y = 1$ whereas the curve generated using the simulation method stops short of this value as the modulator becomes unstable. The difference occurs because the evaluation of σ_e^2 using the PDF method gives no information about modulator stability.

In figure 3.3 the values of K_n are also plotted against m_y. In the PDF method, two values of K_n are obtained by solving equation 3.30 (details of the solution are given in appendix B.2). The upper curve approximately follows the value of K_n obtained by the time-average method, with a fairly constant error. An explanation

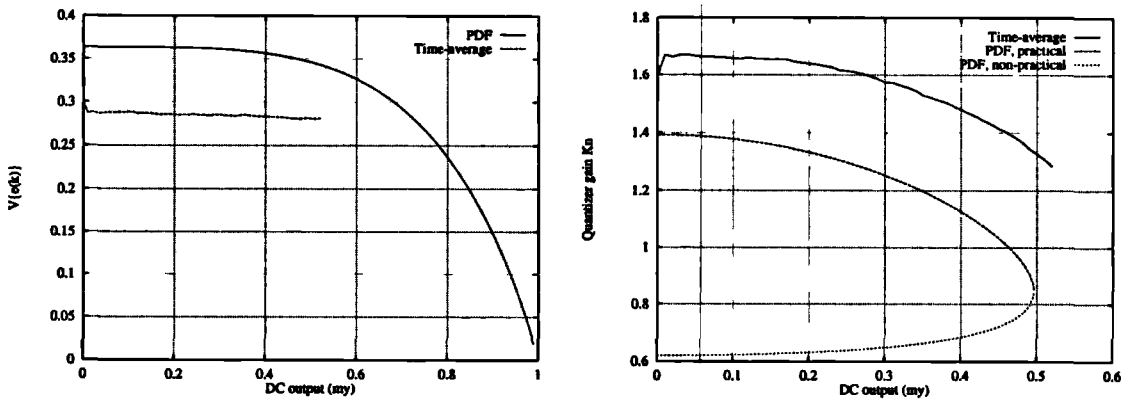


Figure 3.3: σ_e^2 $(= V\{e(k)\})$ and K_n against m_y for modulator {64, 4, 3.5} using PDF and Time-Average Methods

for the lower curve will be given in section 3.5. The two solutions converge at a DC value which is near the maximum stable amplitude (MSA) of the modulator. The value of K_n decreases with m_y and this is associated with an increase in the variance of the quantizer input $n(k)$.

The baseband noise power is plotted for the PDF and time-average methods in figure 3.4, together with a reference curve generated using a direct simulation combined with spectral estimation (refer to appendix A.1). This allows the accuracy of the noise prediction to be evaluated. Note that in the case of the PDF method, the upper $K_n(m_y)$ curve is used in calculating the noise performance, as this curve is followed in practice. Comparing the results, it can be seen that the time-average method produces a fairly good estimate of the modulator noise power, whereas the PDF method has a larger error, although the general trend of the curve is followed.

3.5 Stability Approximation

In this section it will be shown how the quasi-linear model may be used to approximately predict the stability of $\Sigma\Delta$ modulators. The model has been used to analyse stability in [Ard87], using a Nyquist plot in conjunction with the quasi-linear gain. Another related technique is the root-locus approach of [Sti88b], where the pole locations on the unit circle are estimated from the quantizer gain. An ambiguity with these techniques is that the method used to define the quantizer gain obviously affects the results obtained and it is unclear exactly how a gain expressed in

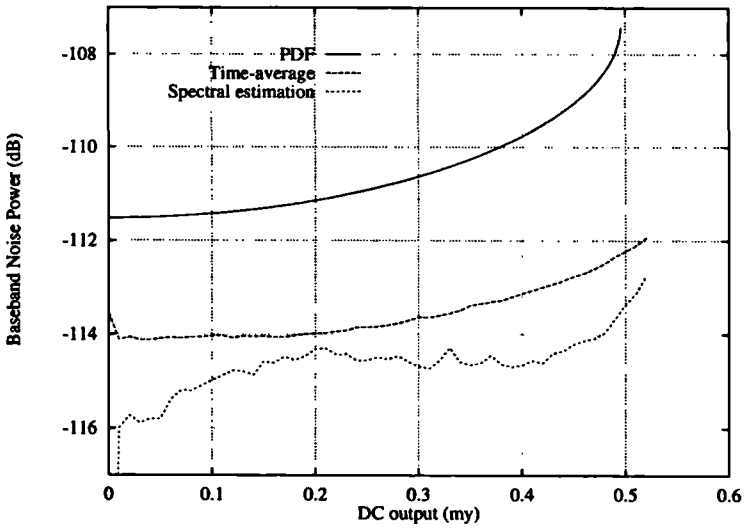


Figure 3.4: Baseband Noise Power (dB) against m_y for for modulator {64, 4, 3.5} using PDF, Time-Average and FFT-measurement Methods

statistical terms relates to the gain which defines the stability of a linear feedback system. These methods are useful, however, in describing the stability *behaviour* of the modulator i.e. they predict that modulator instability is related to a large quantizer input variance, which may occur for large signal levels or NTFs with large power gains.

The stability analysis used in this section uses a different approach which avoids the uncertainty in the choice of quasi-linear gain. The same *results* are obtained as the Gaussian Stability Criterion of [Ris94b], though the method and interpretation are different. The basic idea is that equation 3.30 may be used in conjunction with the DC Coding Property 2.1 to describe a functional modulator which generates a DC output approximating to its input. It will be shown that the region where no solution exists corresponds to an operating region where the modulator is non-functional, and that this can be interpreted as instability.

3.5.1 NTF Power Gain and the Quasi-Linear Model

The existence of a solution to equation 3.30 can be evaluated by considering the Constant Output Power Property (property 2.2). This shows that a balance exists between signal and noise power at the modulator output. By means of the white noise assumption this balance can be expressed as:

$$m_y^2 + \sigma_e^2 P_n(K_n) = \Delta^2 \tag{3.38}$$

where $P_n(K_n)$ is the power gain of $NTF_K(z)$ given by equation 2.13 or 2.14 and σ_e^2 is the quantizer error variance given by equation 3.22.

In equation 3.38, the total noise power $\sigma_e^2 P_n(K_n)$ consists of the quantizer error power amplified by the power gain of the K_n-corrected noise transfer function. As m_y increases, $\sigma_e^2 P_n(K_n)$ must reduce to accommodate the signal power. The equation may be re-expressed as:

$$\frac{\Delta^2 - m_y^2}{\sigma_e^2} = P_n(K_n) \tag{3.39}$$

The characteristic $(\Delta^2 - m_y^2)/\sigma_e^2$ is plotted in figure 3.5 for $\Delta = 1$ and an assumed Gaussian quantizer input characteristic. The curve shows that if equality is to be maintained in equation 3.39, $P_n(K_n)$ must decrease monotonically with m_y. The $P_n(K_n)$ curve is crucial in determining stability. The NTF Scaling Constraint 2.1 defines a lower bound on the value of $P_n(K_n)$ for $K_n = 0$, i.e. by equations 2.4 and 2.14 $P_n(K_n) = 1$ for $K_n = 0$. Therefore if K_n reduces to zero as m_y approaches unity (assuming $\Delta = 1$) it is theoretically possible for the maximum output level $m_y = 1$ to be sustained. More detailed properties of the $P_n(K_n)$ curve have been investigated in [Ris94b] and three distinct curves have been identified. These are defined below.

1. Type I: $P_n(K_n)$ is increasing and $min\{P_n(K_n)\} = P_n(0) = 1$
   characterises non-chaotic first and second order NTF with distinct zeros, e.g. first order Tewksbury filter [Tew78], $NTF(z) = 1 - z^{-1}$

2. Type II: $P_n(K_n)$ is increasing and $min\{P_n(K_n)\} = P_n(0^+) > 1$
   characterises non-chaotic second order NTF with double DC zeros, e.g. 2nd order Tewksbury filter $NTF(z) = (1 - z^{-1})^2$

3. Type III: $P_n(K_n)$ is U-convex
   characterises all chaotic and high order NTFs. e.g. third order Butterworth NTF, $P_n = 3.5dB$

The characteristic $P_n(K_n)$ curves for the three examples are shown in figure 3.6. In the type II and III curves there is a sharp discontinuity at $K = 0$ since $P_n(0) = 1$

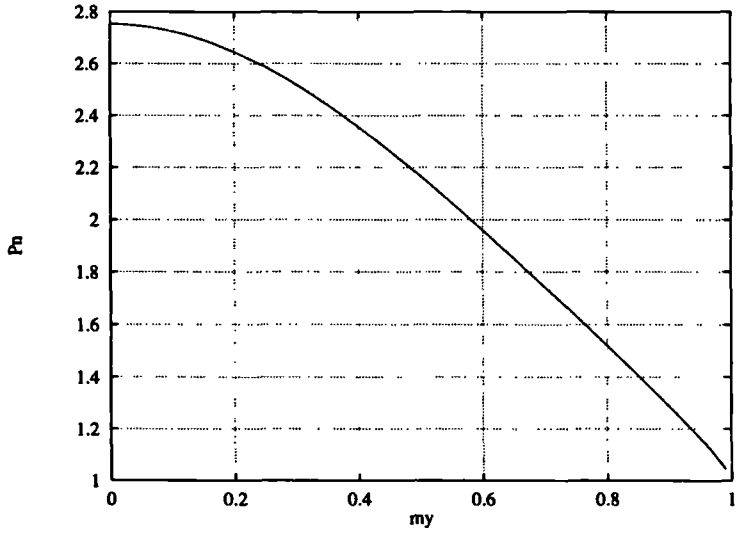


Figure 3.5: $P_n(m_y)$ characteristic for a Gaussian quantizer input, found by evaluating the left hand side of equation 3.39

in all cases. Therefore in the above definition of the type II curve, the notation $min\{P_n(K_n)\} = P_n(0^+) > 1$ is used to indicate that the characteristic only applies for $K_n > 0$.

The stability of modulators employing NTFs with one of the three characteristic curves can be determined by comparing figures 3.5 and 3.6, as described below.

Stability of Modulators with Type I NTFs

For type I curves, as m_y increases, equality can be maintained in equation 3.38 by a monotonic decrease of K_n. There is a value of K_n which allows equation 3.38 to be solved for any m_y in the range ± 1. Accordingly, there is also a solution to equation 3.30 for any m_y.

Stability of Modulators with Type II NTFs

For the type II curve, K_n will only decrease monotonically until $P_n = P_n(0^+)$. In figure 3.7, the solution of equation 3.30 is plotted against m_y for the type I and II curves and $\Delta = 1$. It can be seen that for the type II curve, a solution for K_n exists up to a maximum input level. The region of the curve above $m_y = 0.8$ corresponds to the discontinuous region in the $P_n(K_n)$ characteristic. Therefore it is also theoretically possible for a modulator with a type II NTF to achieve maximum input level.

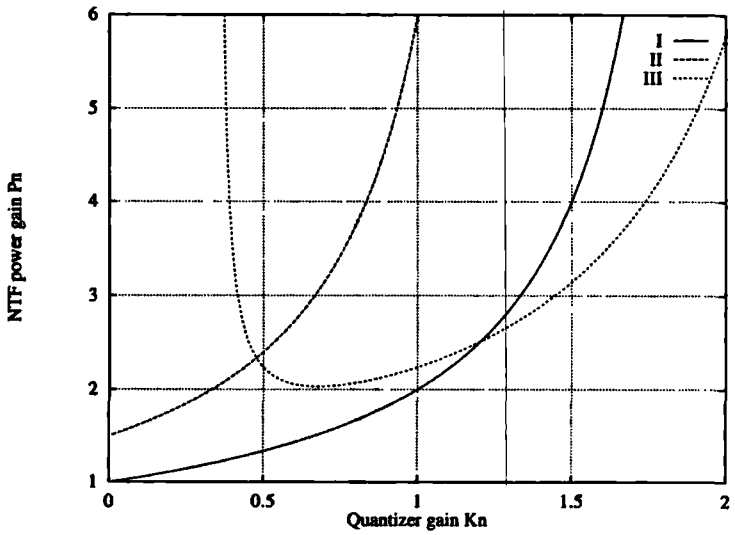


Figure 3.6: $P_n(K_n)$ curves for example type I, II and II NTFs

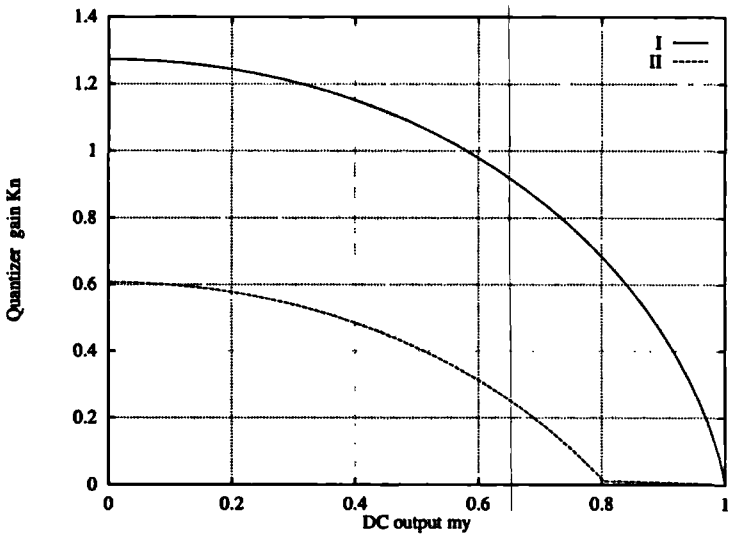


Figure 3.7: Gaussian $K_n(m_y)$ curves for example type I, II NTFs

Stability of Modulators with Type III NTFs

The characteristic curve of the type III modulator has a minimum $min\{P_n(K_n)\} > 1$. As m_y increases, K_n decreases towards this minimum, causing $P_n(K_n)$ to decrease in order to maintain equality in equation 3.38. Above the minimum, there are two possible values of K_n for every $P_n(K_n)$. These values correspond to the upper and lower branches of the $K_n(m_y)$ characteristic, as shown in the example of figure 3.3. Of the two $K_n(m_y)$ curves, the upper branch is followed for real simulations, corresponding to the positive slope of $P_n(K_n)$. This curve represents a stable *equilibrium* in the sense that if K_n temporarily increases due to short-term perturbations in the signals, $P_n(K_n)$ increases, causing the noise power circulating in the loop to increase. This causes K_n to reduce back to its equilibrium value [Ris94b]. Conversely the lower $K_n(m_y)$ branch represents an unstable equilibrium which is not observed in real modulators. This also explains why $K_n = 0$ is never reached. Although it is a theoretically possible solution, the modulator would need to pass through a region of unstable equilibrium to reach this point.

The minimum on the $P_n(K_n)$ curve (figure 3.6) corresponds to a value of m_y on the $P_n(m_y)$ curve (figure 3.5). For any further increase in m_y , P_n cannot reduce further and there is *no solution* to equation 3.30. This means that for an input level m_x exceeding this value of m_y, $m_x \neq m_y$ and the modulator no longer codes the DC input level. By the DC Coding property, the modulator is then unstable. This has been demonstrated by example in section 2.3.3.

Identifying the $min\{P_n(K_n)\}$ point allows the maximum stable amplitude of the modulator to be predicted, therefore the above interpretations provide a link between the stability of the modulator and the characteristics of its NTF. An interesting outcome of the analysis is that the stability of high order modulators is independent of loop order - it is only the $min\{P_n(K_n)\}$ in the power gain curve that is important.

The two assumptions made in here are that the AC component of quantizer input signal has a Gaussian characteristic, and the additive error has white spectral characteristics.

The existence of a solution for K_n for a given m_y means that a solution also exists for the variance of the quantizer input σ_n^2. This implies that the quantizer input is bounded and the modulator is stable. Where a solution exists (type I and II curves), the case where $m_y = 1$ is special in the sense that $\sigma_e^2 = 0$, therefore by equation 3.28, $\sigma_n^2 = 0$ and there is no noise circulating in the loop. The zero noise

condition allows an output saturated to full scale to exist (under the approximations of the model). However, under this condition the noise shaping of the modulator breaks down. For a time-varying input such as a sinewave, it is clear that the noise shaping characteristics must be maintained for baseband error cancellation to occur. Therefore the signal power cannot reach unity whilst modulator functionality is maintained.

3.6 Investigations and Results

3.6.1 Noise Model

Detailed results are now presented for the estimation of the quasi-linear parameters and baseband noise power using both the PDF and time-average methods described above. For these results, we concentrate on higher order modulators of order four and six with oversampling ratio $L = 64$. For each order, four NTFs of power gains A($P_n = 2.5dB$), B($P_n = 3dB$), C($P_n = 3.5dB$) and D($P_n = 4dB$) have been modelled using the PDF and time-average methods.

In figure 3.8 and 3.9, σ_e^2 is plotted against m_y for both methods with modulator orders $N = 4$ and $N = 6$. It has been predicted that, for an assumed Gaussian input, the values of σ_e^2 depend only on the DC input and not the NTF parameters. The values obtained using the time-average method for different P_n follow similar though not identical curves to each other and there is an error between these curves and the theoretical Gaussian curve. The differences are attributed to errors in the Gaussian assumption. As an example of this, the PDF of the quantizer input has been determined for the fourth order modulator with NTF C with an input $m_x = 0.3$ and power gain $P_n = 3.5$. This curve are plotted in figure 3.10, together with the theoretical Normal curve (i.e. the Gaussian curve with values of mean and variance observed during the simulation). Also plotted is the PDF for the fourth order modulator with NTF A with a DC input $m_x = 1$. In this case the quantizer input has become dominated by a limit cycle and certain PDF bins are periodically visited. This explains the irregularity of the $\sigma_e^2(m_y)$ characteristic for this modulator.

The values of K_n are plotted against m_y for both PDF and time-average methods in figures 3.11 and 3.12. The curves for the PDF method have been obtained by solving equation 3.30 using the method described in appendix B.2. As described in section 3.5.1, the upper branch of the PDF method K_n curve represents the solution

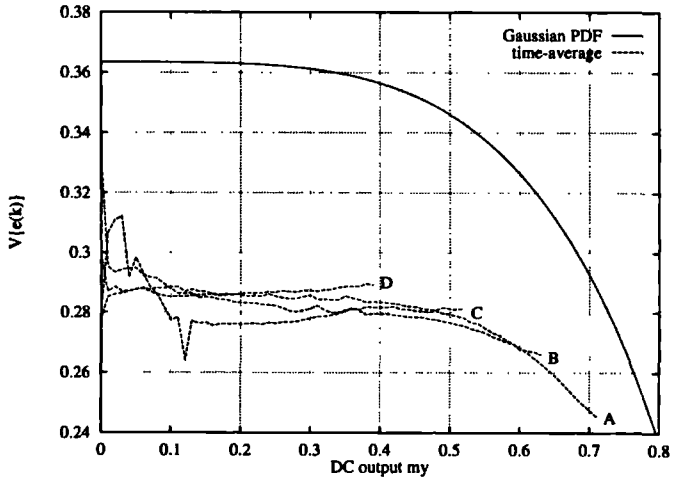


Figure 3.8: PDF and time-average σ_e^2(=V{e(k)}) against m_y for $N = 4$ with filters A, B, C, D

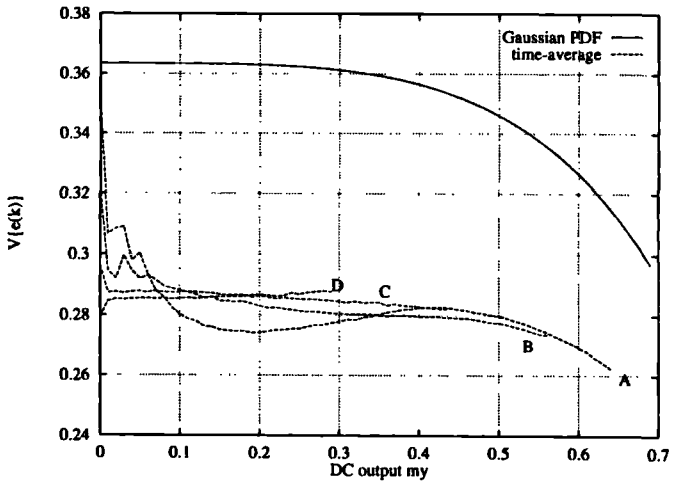


Figure 3.9: PDF and time-average σ_e^2(=V{e(k)}) against m_y for $N = 6$ and filters A, B, C, D

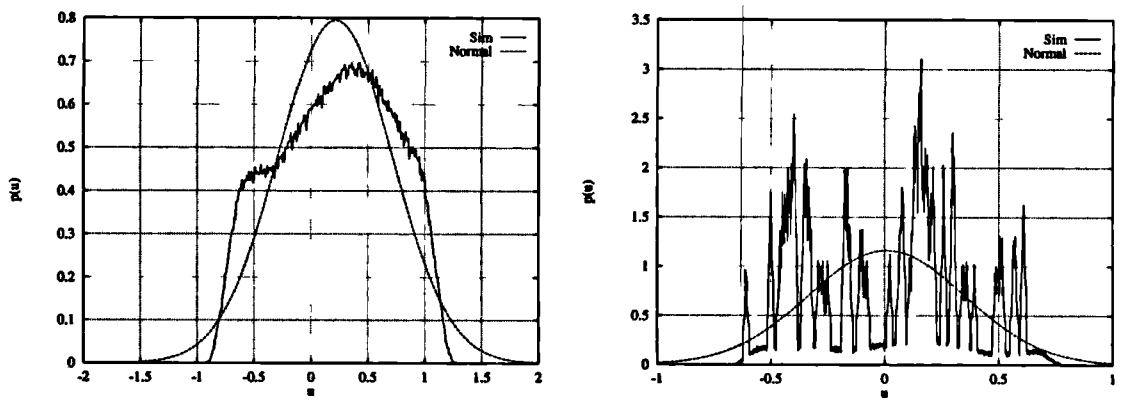


Figure 3.10: PDF of quantizer input $u(k)$ for (a) $N = 4$, $P_n = 3.5$, $m_x = 0.3$ and (b) $N = 4$, $P_n = 2.5$, $m_x = 0.1$

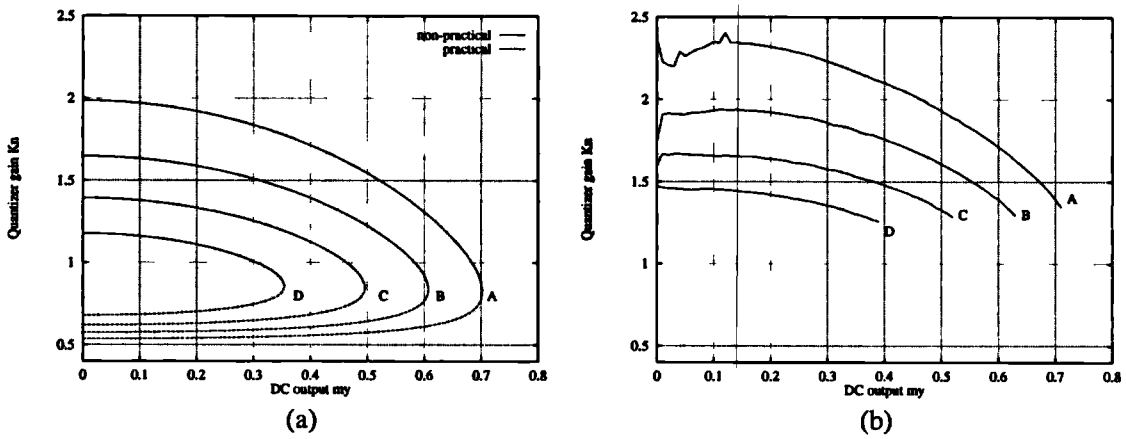


Figure 3.11: K_n against m_y for $N = 4$ and filters A, B, C, D with (a) PDF method and (b) time-average method

obtained for a practical stable modulator.

There is a difference between the results obtained using the two methods, with the time-average method producing larger values of K_n. The errors in the PDF method are due to two assumptions used in obtaining the solution to equation. 3.30. Firstly the Gaussian assumption in used in obtaining the value of σ_e^2 and in evaluating equation 3.23. Secondly, the white noise assumption is used in evaluating equation 3.28. In contrast, neither assumption is used in evaluating K_n with the time-average method, therefore this curve may be used as a reference. Despite the assumptions, both sets of curves follow the same trends, with a fairly constant offset.

Finally the baseband noise power is plotted for both the PDF and time-average

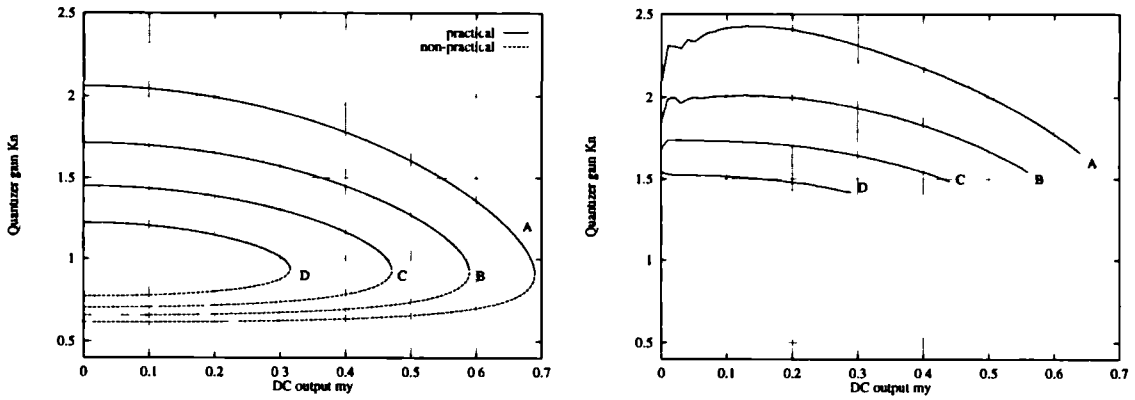


Figure 3.12: K_n against m_y for $N = 6$ and filters A, B, C, D with (a) PDF method and (b) time-average method

methods in figures 3.13 and 3.14. These results show how the errors in $K_n(m_y)$ and σ_e^2 affect the accuracy of the noise prediction. Note that the white noise assumption is used in both methods in the evaluation of equation 3.31. In figure 3.13(b) and 3.14(b), reference curves have also been generated using direct simulation and noise measurement using the spectral estimation approach of appendix A.1. The non-monotonic shape of these curves are caused by idle tones in the modulator, which are not predicted by the quasi-linear model due to the white noise assumption. Comparing the results, it can be seen that the time-average method produces a good estimate of the modulator noise power, with deviations typically in the order of $2 - 3$ dB. The PDF method produces a larger error in the order of $4 - 5$ dB, however, it should be realised that the quasi-linear model does predict the general shape of the noise curves. It will be shown in chapter 4 that the quasi-linear model is more accurate when the inclusion of a dither source is modelled.

3.6.2 Stability of High Order Modulators

In this section results are presented, comparing the predicted stability of high order modulators using the quasi-linear model to the stability measured by direct simulation. To investigate the accuracy of the prediction for a wide range of modulators, the DC MSA of modulators with Butterworth NTFs with $L = 64$ has been compared to the Gaussian prediction over a range of power gains and for orders 3 to 6 in figures 3.15 and 3.16.

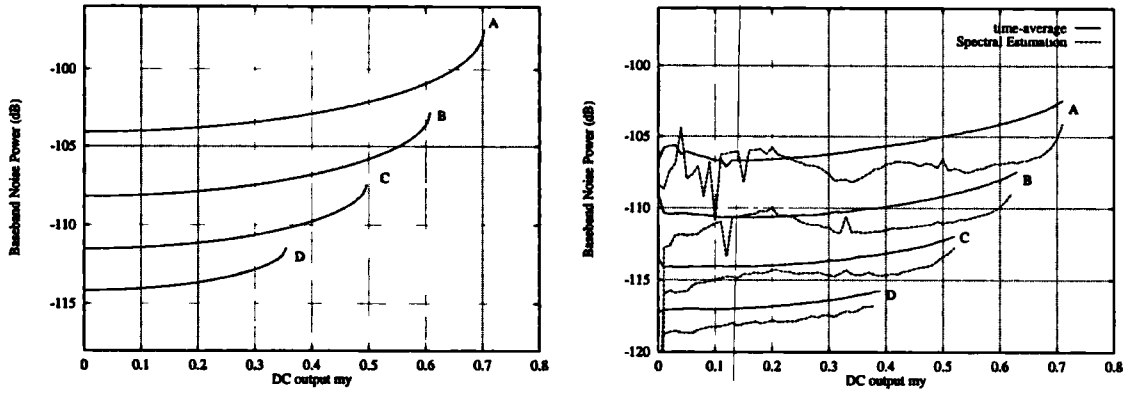


Figure 3.13: Baseband Noise Power (dB) against m_y for $N = 4$ and filters A, B, C, D with (a) PDF method and (b) time-average method

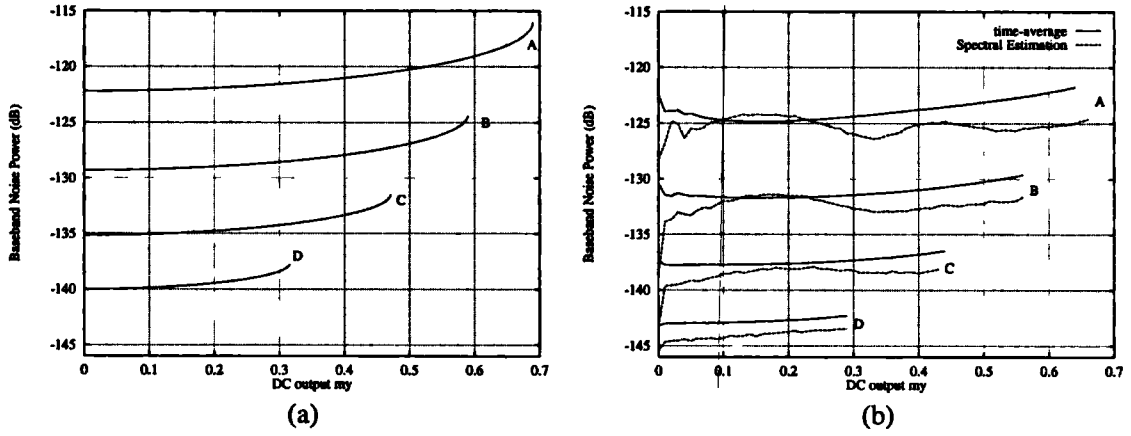


Figure 3.14: Baseband Noise Power (dB) against m_y for $N = 6$ and filters A, B, C, D with (a) PDF method and (b) time-average method

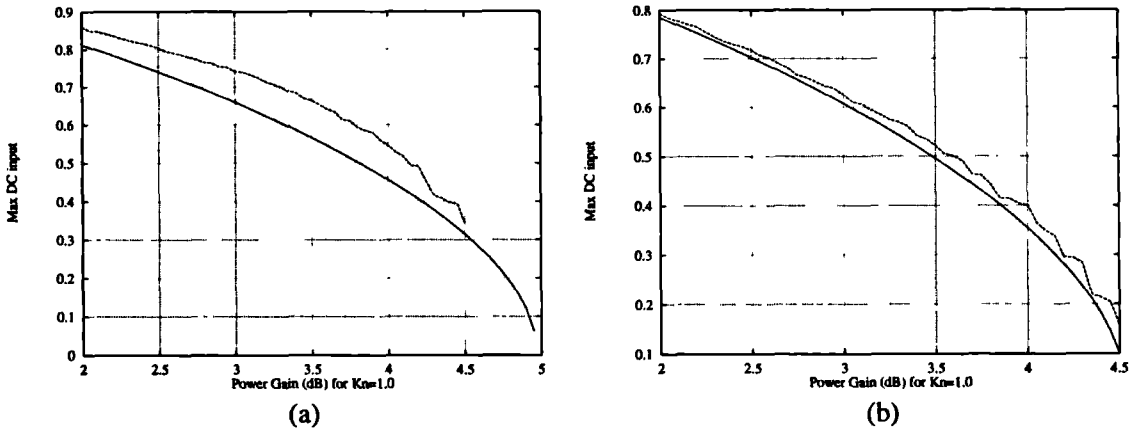


Figure 3.15: Comparison of theoretical and experimental Maximum DC input level, for (a) $N = 3$ and (b) $N = 4$

The MSA has been obtained by repeated simulations and the use of a local search algorithm to 'home-in' on the maximum input amplitude. Considerably reduced simulation times have been obtained by beginning the search at an input level far above the MSA. The modulator becomes unstable in a very short number of simulation samples, allowing the search to proceed quickly. For each test, the modulator is deemed to be stable if the quantizer input is bounded by ± 1000 after 1.5 million samples.

The results show that the shape of the theoretical stability curves is closely followed by the simulated modulators, especially for the fourth and higher order modulators. The absolute stability prediction is accurate, however it is noticed that as the modulator order increases, the simulated MSAs decrease relative to the predicted MSAs (see figures). This leads to an artificially accurate result for the fifth order modulator. Further work is required to establish the reason for this trend and determine how the accuracy of the PDF and white noise assumptions affect the results.

Despite these reservations, the model reveals a strong and encouraging link between the $min\{P_n(K_n)\}$ of the NTF and modulator stability for Butterworth NTFs. The method allows very rapid predictions to be made about modulator stability.

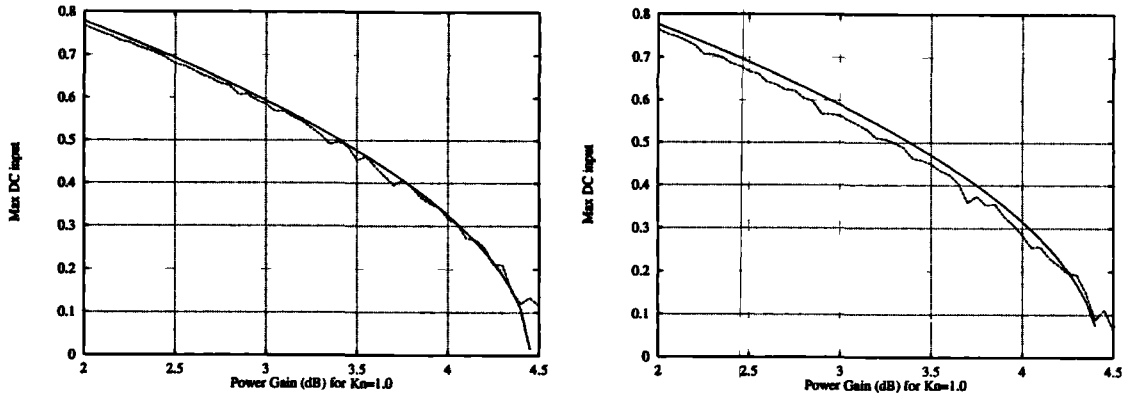


Figure 3.16: Comparison of theoretical and experimental Maximum DC input level, for (a) $N = 5$ and (b) $N = 6$

3.7 Summary

In this chapter the quasi-linear model which was introduced in chapter 2 has been developed. The model is useful in obtaining rapid and reasonably accurate predictions of noise and stability. It relates modulator behaviour to parameters which can be obtained for general NTFs. The noise analysis is based upon the work of Ardalan and Paulos [Ard87], then extended to evaluate the quantizer gain by solving a nonlinear equation. The stability analysis is based upon a novel interpretation of solution of this equation.

Two quantizer gains are defined: the noise gain K_n and signal gain K_s. The latter is defined to help in the formulation of the model, but is eliminated from the final expressions. The gains are defined such that the variance σ_e^2 of the additive quantization error is minimised. The quasi-linear analysis proceeds by either the PDF method or the time-average method. In the former case an assumption is made that the AC component of the quantizer input has a particular probability density function, in this case a Gaussian distribution. This leads to a direct evaluation of σ_e^2 which is independent of the NTF, and a nonlinear equation which can be solved to find K_n. Using the second method, time-averages are derived which allow K_n and σ_e^2 to be measured during a simulation of the modulator.

The key results of the noise analysis are that K_n and σ_e^2 reduce as the input amplitude increases. The reduction in quasi-linear gain causes the magnitude response of the NTF to vary with input level and the baseband noise power to increase. The

modelling errors are fairly small and constant with input amplitude.

The stability of the modulator is also described in terms of the solution of the exsitence of a solution to a nonlinear equation which defines K_n. A solution will only exist if the signal and noise power can be accommodated under the constraints of the Constant Output Power Property. The variation in NTF noise power gain with K_n is crucial to this relationship, and it has been shown the minimum power gain defines the MSA of the modulator. The model predicts stability up to full scale with first and second order Tewksbury modulators and conditional stability in higher order modulators.

Simulation results show that the estimates of MSA obtained using this approach are very accurate, both in qualitatively and quantitatively. It is noted that the accuracy is dependent on modulator order and further work is required to ascertain how the accuracy relates the the fundamental PDF and white noise assumptions.

The model will prove useful for explaining modulator behaviour in subsequent chapters and the techniques will be developed in chapter 4 for modelling the noise performance of dithered and bit-flipping modulators using the PDF method; and in chapter 6 for estimating the average pulse repetition frequency of a class of bit-flipping modulators using the time-average method.

Chapter 4

Linearisation of Sigma Delta Modulators using Bit-Flipping

4.1 Introduction

In chapter 2 it has been shown that limit cycles occur in the state space and output of the $\Sigma\Delta$ modulator due to the nonlinear quantizer in the loop. The resulting periodicity in the bitstream causes tonal components to appear in the noise spectrum. The tones occur predominantly for rational DC inputs, since it is possible to represent these inputs by the average level of a periodic output sequence. The tonal components are responsible also for noise modulation, where the baseband noise power varies with changes in the DC input. These artifacts can occur in both high and low order modulators, though the problem becomes less severe as the modulator order increases. Although the occurrence of both tonal components and noise modulation has been identified with DC inputs, more complex band-limited inputs will also give rise to nonlinear artifacts, as they can be represented as a slowly varying DC input. Furthermore, for A-D converters, any DC drift in the modulator may cause the character of the nonlinearity to become time-dependent.

In the context of an audio system, the quality of the conversion can be seriously degraded, especially since the ear is sensitive to tonal components and noise modulation.

The purpose of this chapter is to investigate the linearisation of $\Sigma\Delta$ modulation using schemes based upon dither, and to introduce a new linearisation scheme that uses the technique of bit-flipping introduced in chapter 2.

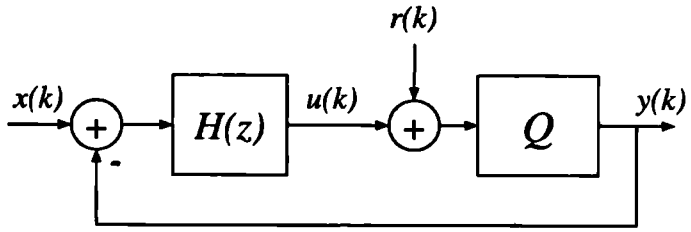


Figure 4.1: Block Diagram of Dithered ΣΔ modulator

Section 2.4.6 introduced the principle of linearisation using dither. Dither is a random or pseudo-random noise source which is added to the quantizer input signal (figure 4.1). Vanderkooy and Lipshitz [Van89] have shown that for *multibit open-loop quantization*, triangular PDF dither spanning twice the quantizer interval will linearise the time-averaged quantizer transfer function and ensure zero noise modulation. The linearisation is at the expense of a dynamic range penalty. Smaller penalties can be obtained using triangular PDF high-pass dither for oversampled systems, because the dither noise power falling into the baseband is lower.

Although there has been considerable research into the dithering of one-bit ΣΔ modulation, much of this has been based upon the 'blind' application of the above work [Nor93]. As a result, the conventional techniques of linearisation using dither with a prescribed PDF and frequency response, although successful, are not necessarily optimal in terms of hardware efficiency and performance. In this chapter, we show that it is possible to linearise higher order modulators without the introduction of a random component. Consequently, it is possible to achieve enhanced conversion linearity with a lower hardware cost that conventional dither. Furthermore, under certain conditions it is possible to linearise the modulator with a lower dynamic range penalty than conventional dithering.

The nonlinear behaviour of higher order modulators is not well understood and with the exception of the work on chaotic modulators by Motamed [Mot96] and the generalisations of Ledzius and Irwin [Led3], there has been very little published work on the relationship between the spectral characteristics of the modulator output and the time-domain limit cycles present. The only rigorous study into the spectrum of dithered modulators has been for the first order modulator [Cho91]. Consequently, in the work presented in this chapter, the tonal characteristics of different dithering schemes with high order modulators have been obtained directly by simulation. However, analysis has been used where possible, and in particular, the technique of

quasi-linear modelling has been used to identify a theoretical relationship between rectangular PDF dither and the deterministic bit-flipping scheme which will be introduced later.

4.2 Modelling Dither as Bit-flipping

We begin with a view of dithering which leads to a clearer understanding of the mechanisms which are responsible for tone attenuation and the effect of dither on modulator behaviour. The perspective is summarised by the following proposition, which relates to the injection of a dither signal at the input of the quantizer (figure 4.1):

Proposition 4.1 *The effect of any dither input $r(k)$ is to change the decision made by the quantizer for some samples.*

The decision will change whenever the dither signal causes the sign of the quantizer input $u(k)$ to change. The conditions that are required for this to occur are:

$$|r(k)| > |u(k)| \tag{4.1}$$

$$sgn\{r(k)\} \neq sgn\{u(k)\} \tag{4.2}$$

Due to this selective change in quantizer state, a dithered quantizer may be modelled as a bit-flipping operation on an undithered quantizer. This concept is illustrated in figure 4.2. In the equivalent model, the bit-flipping operator is triggered whenever conditions 4.1 and 4.2 are jointly satisfied. It will be demonstrated in sections 4.3.1 and 4.5 that the bit-flipping has the effect of disrupting the limit cycles in the modulator output, causing periodic components of the quantizer error to become attenuated and resulting in an increase in modulator linearity.

4.2.1 Interpretation of Bit-flipping Model

In this section we discuss the inplications of th bit-flipping model and establish whether it is possible to define an optimal dither specification, in terms of dither amplitude, probability density function and frequency response, which applies to a general $\Sigma\Delta$ modulator.

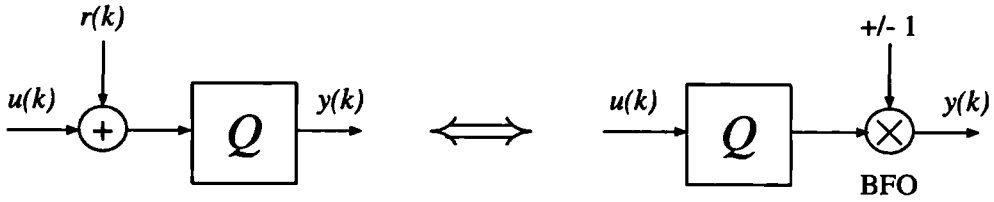


Figure 4.2: Dither as an equivalent bit-flipping operation

Dither Amplitude

Condition 4.1 shows that the probability of bit-flipping occurring depends on the instantaneous dither and filter output amplitudes. The rate of bit-flipping (i.e. the percentage of samples for which bit-flipping occurs) will increase as the peak dither amplitude is increased. This causes the limit cycles to become more disrupted, and therefore the tone amplitudes decrease with dither amplitude.

The peak dither amplitude required to excite a given bit-flipping activity depends on the PDFs of $r(k)$ and $u(k)$, and the relative scaling of $u(k)$. Due to the scaling invariance property (i.e. scaling $u(k)$ by a positive constant does not affect the operation of the modulator), the scaling of $u(k)$ is arbitrary , therefore unlike multibit quantizers, it is not possible to specify an optimal dither magnitude which generalises to all modulators. Furthermore, the rate of bit-flipping required to linearise the modulator is dependent on the persistence of the limit cycles, which is related to the NTF characteristics, for example the order of the modulator [Cha90] and the zero placement. This view of dithering is also useful in explaining why a high level of dither is required to attenuate the high frequency tones. The tone near $Lf_s/2$ is associated with a persistent alternating one-zero pattern [Ris94b]. Frequent bit-flipping is required to break up this pattern, therefore the dither amplitude must be high enough to cause condition 4.1 to be frequently satisfied.

Dither Probability Density Function

As with the dither amplitude, the importance of the dither PDF essentially relates to the bit-flipping rate resulting from the satisfaction of condition 4.1. To achieve the same bit-flipping rate as rectangular PDF dither, triangular PDF dither requires a greater peak dither amplitude, due to the smaller probability of large dither magnitudes occurring. This result explains the simulation results in [Dun94], which show that the amplitude of triangular PDF dither required to linearise the modu-

lator is greater than rectangular PDF dither, but in terms of linearity there is no distinction between the two PDFs. The relative unimportance of the dither PDF is exemplified by the successful application of one-bit dither to the linearisation of high order modulators [Gal93, Dun95].

One bit dither is essentially a sampled pseudo-random noise source quantized to a single bit i.e. it is a two level sequence which randomly oscillates between the two levels. The PDF of one bit dither comprises two equal weighted impulses at plus and minus the dither amplitude. It is attractive from an implementation perspective because it can be generated using a maximum length sequence [Mut96]; and for A-D conversion, it can be converted to an analogue dither source using only a sample-and-hold.

It will be shown in section 4.3.3 that although the dither PDF is unimportant for sucessful linearisation, it is critical with regards to the stability of the modulator.

Dither Frequency Response and Noise Penalty

Equation 2.25 relates the noise spectrum at the output of the modulator to the dither source $R(z)$ and quantizer error $E(z)$. Since the dither source modifies the operation of the quantizer, the quantizer error $E(z)$ is dependent on the dither signal $R(z)$. Unlike the application of dither to multibit quantizers, it is not possible to apply the principle of superposition in equation 2.25. Therefore to calculate the baseband noise power of a dithered modulator, it is not appropriate to simply add the baseband dither power to the baseband noise power of an identical undithered modulator. Any dither source which is uncorrelated with $u(k)$ with cause an increase in the variance at the input to the quantizer, *regardless of its spectral properties.* In terms of the quasi-linear model, this causes the quantizer gain K_n to reduce and the error variance σ_e^2 to increase [Ris94b], therefore even a dither source with only out-of-band noise energy will cause the baseband noise power to increase. A quasi-linear model for dithered $\Sigma\Delta$ modulator will be rigorously defined in section 4.4.

4.3 Implementing Dither with Bit-Flipping

In the previous section it has been argued that in terms of linearisation, the effectiveness of different dithering schemes essentially relates to how well the limit cycles in the one-bit output are broken up by the resulting bit-flipping.

A pivotal step which develops from proposition 4.1 is the possibility of implementing dither directly as a bit-flipping operator (BFO), which can be used to directly break up limit cycles. A practical advantage of this scheme (for A-D converters) is that the linearisation algorithm can be implemented in the digital-domain and therefore avoid the problems of implementing an analogue noise source (refer to section 2.4.6). A further possibility is for the BFO control algorithm to observe the limit cycles which are to be eradicated, leading to a more intelligent and efficient linearisation scheme.

4.3.1 Bit-flipping using Limit Cycle Detection

A starting point is to consider the case of using bit-flipping to linearise a modulator which generates known limit cycles. We consider an example modulator with parameters {64, 4, 4.0} and a DC input $m_x = 1/2^{20}$. This modulator generates a fairly random quantizer error for the first 500,000 samples (figure 4.3). Afterwards the modulator becomes attracted to a limit cycle with a period of 44 samples. Figure 4.4(a) and (b) shows the spectra after an initial offset of 600,000 and 750,000 samples, respectively. The first plot shows the limit cycles developing, and in the latter plot, the modulator output has become purely tonal.

Note that in figure 4.4, as the limit cycles are developing, the behaviour of the modulator is time-varying, and due to the FFT averaging, the tonal components become superimposed upon the noise floor. This is responsible for the existence of both noise and tonal components in figure 4.4(a).

We now assess whether bit-flipping can be used to eliminate the idle tones by preventing the modulator becoming attracted to the period 44 limit cycle. An identical modulator has been simulated, with the inclusion of a single bit-flip forced at sample number 800,000. Figure 4.5(a) and (b) shows the wideband spectra of the modulator output, measured after an initial offset of 800,000 and 900,000 samples, respectively. After 800,000 samples the limit cycle has been disrupted and the quantization noise is randomized. The limit cycle re-develops between 850,000 and 900,000 samples. This result suggests that flipping as irregularly as every 1/50,000 samples would be sufficient to prevent purely tonal behaviour, for this particular modulator and input.

It is apparent from these results that it is possible to eliminate certain facets of nonlinear behaviour using bit-flipping, as long as sufficient knowledge is obtained of

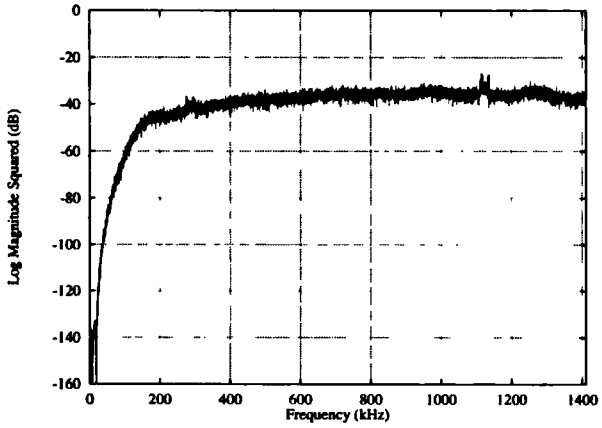


Figure 4.3: Baseband spectrum of modulator {64, 4, 4.0} with $m_x = 1/2^{20}$ after an offset of 500,000 samples.

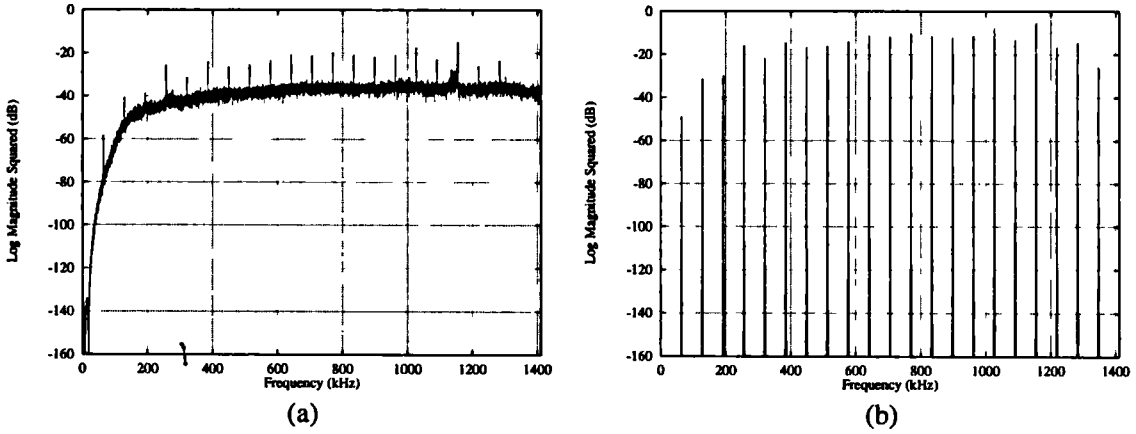


Figure 4.4: Baseband spectrum of modulator {64, 4, 4.0} with $m_x = 1/2^{20}$ after an offset of (a) 600,000 samples and (b) 750,000 samples.

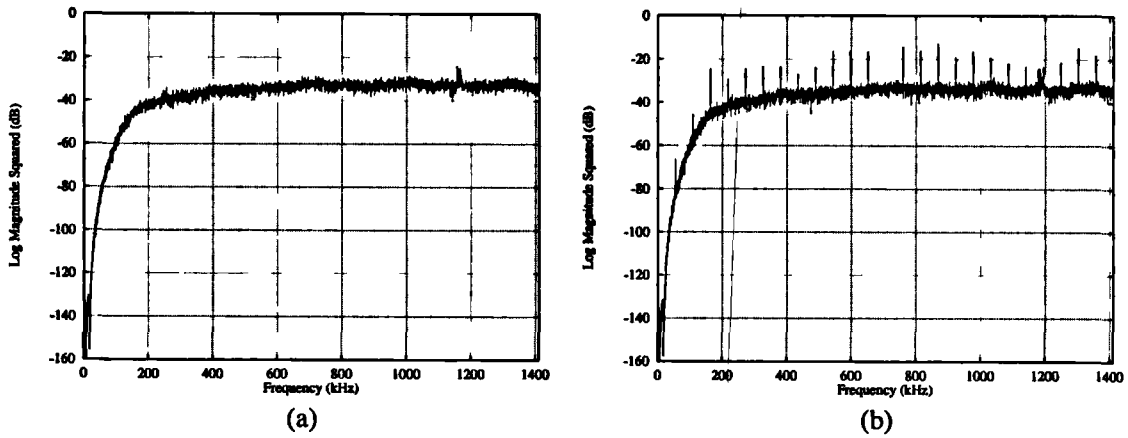


Figure 4.5: Wideband spectrum of undithered modulator with parameters {64, 4, 4.0} and $m_x = 1/2^{20}$ after a single bit-flip at 800,000 samples and offset of (a) 800,000 samples and (b) 900,000 samples

the limit cycles which need to be eradicated. A possible scheme for more general linearisation uses a limit cycle detector (LCD), as shown in figure 4.6 to establish when particular limit cycles occur and then appropriately trigger the BFO. This approach has been investigated in the author's papers [Mag95a, Mag95b] and a reprint of the second paper is presented in Appendix C.1. To summarise, the LCD detects sequences of alternating patterns such as {1, 1, -1, -1, 1, 1, -1, -1, ...} and {1, -1, 1, -1, 1, -1, 1, -1, ...} which indicate the presence of components near $Lf_s/2$. Under a predetermined set of conditions, which include a random condition, this pattern is broken up by triggering the BFO. Simulation results presented in appendix C.1 indicate that randomly terminating the one-zero pattern has the effect of attenuating tones in the baseband. A possible explanation proposed in the paper is that the phase of the low frequency components becomes randomized by the bit-flipping.

Although this technique is successful in concept, and has the advantage of digital-domain implementation, there are two significant disadvantages, which are briefly described in the paper:

1. For some DC inputs, the modulator becomes 'trapped' by highly attracting limit cycles which are not detected by the LCD, and under these conditions the linearisation fails. This is observed in the noise modulation plots as localised regions where the baseband noise power falls dramatically.

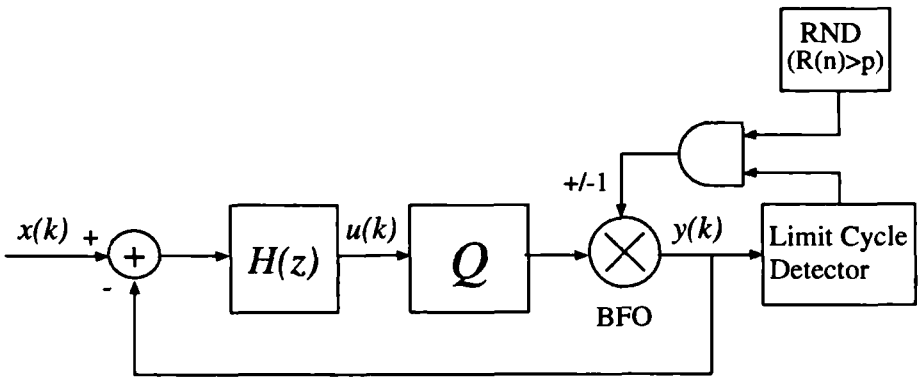


Figure 4.6: Linearisation using a Limit Cycle Detection

2. Unstable regions have been found to occur, for inputs considerably lower that the maximum input level. The instability becomes worse as the BFO activity increases, and consequently it is not possible to provide sufficient bit-flipping to eliminate the persistent high frequency tones.

Due to these problems, the limit cycle detection technique is not pursued here and an alternative dither emulation technique is investigated.

4.3.2 Dither Emulation Using Bit-flipping

Acknowledging the conceptual bit-flipping action of dither leads to the possibility of *emulating* dither by approximately mapping the dither onto an equivalent bit-flipping operation. Referring to the dither definitions in figure 4.7 and condition 4.1, a necessary condition for bit-flipping to occur at sample k is $|u(k)| < B$. This is not a sufficient condition, however, since the dither has a random element which means that bit-flipping will not occur on all samples which satisfy this condition. To obtain an equivalent bit-flipping operation to dither, the fundamental hypothesis is now made that linearisation can be achieved without requiring a random element. A way of achieving this is to simply remove the random element and use the following bit-flipping condition:

Condition 4.1 (Bit-Flipping Condition) *Invert quantizer state if:*

$$|u(k)| < B \tag{4.3}$$

where B is a system constant, termed the quantizer input bound. The technique is termed *deterministic bit-flipping* (DBF).

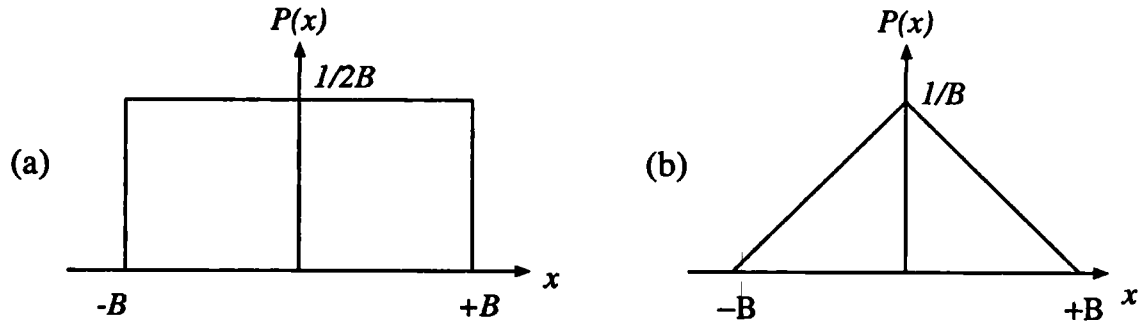


Figure 4.7: Definition of Dither Parameters: (a) Rectangular PDF, (b) Triangular PDF.

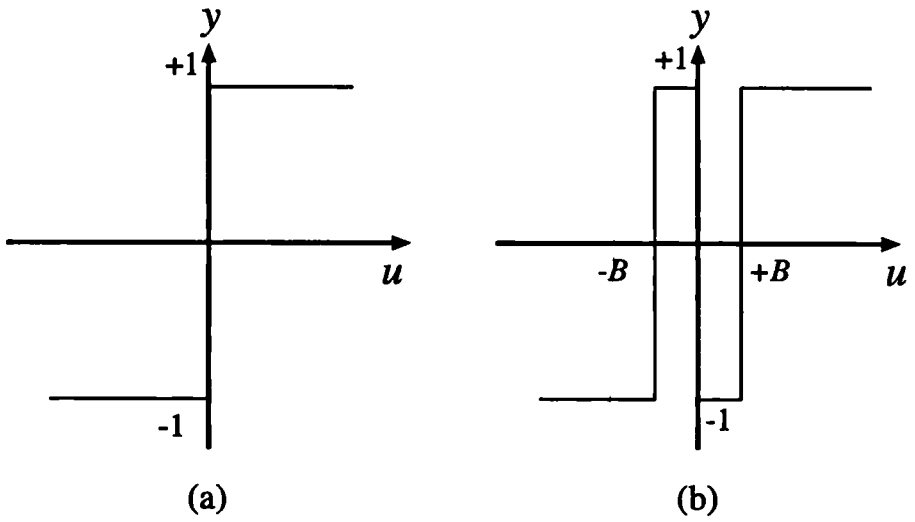


Figure 4.8: Quantizer Transfer Function: (a) Standard Quantizer, (b) DBF Quantizer PDF.

Due to the *deterministic* nature of the bit-flipping, it is possible to map the bit-flipping operation directly onto the quantizer. The effect of bit-flipping is therefore to modify the transfer function of the quantizer nonlinearity, as illustrated in figure 4.8. We conjecture that by increasing the complexity of the quantizer error, the modified quantizer tends to cause more complex limit cycles with lower tonal content.

Evidence that the modified quantizer does enhance the linearity with an appropriate choice of B is demonstrated in the following example.

Linearity of Deterministic Bit-Flipping

To demonstrate the effectiveness of DBF, simulations have been performed of the DBF system using an example modulator with parameters {64, 4, 3.0}. In figures 4.9 and 4.10 the modulator spectrum is plotted for a DC input of 1/512 and values: $B = 0$, $B = 0.04$. $B = 0.07$. Over the range of simulations shown, increasing B results in a reduction in the amplitudes of the baseband tones, indicating that DBF linearises the modulator.

A summary of the linearity of different modulators orders using DBF has been obtained by evaluating the tone amplitudes of the dominant baseband tone f_l and its harmonics; and the high frequency tone f_h, with an input $m_x = 1/512$. The measured frequencies are: 5512.5 Hz, 11025.0 Hz, 16537.5 and 1408.4 kHz. The tone amplitudes have been measured relative to the noise floor using the method of appendix A.3. In figures 4.11 and 4.12 the maximum of the amplitudes of the baseband tones, and the amplitude of the high frequency tone are plotted against B for orders 3 to 6, and parameters $L = 64$, $P_n = 3.0$. These results give a 'snapshot' of the linearity of the modulator.

For fourth and higher order modulators, high linearity occurs periodically with B i.e. there are several values of B which result in good tone attenuation. The greatest attenuation of the high frequency tone occurs close to the second minima of the maximum low frequency tone amplitude, therefore it is possible to attenuate both the high and low frequency tones with an appropriate choice of B.

For the third order modulator, the baseband attenuation is less effective and it is not possible to attenuate both low and high frequency tones. This is possibly due to the lower complexity limit cycle structure associated with low order modulators.

Further results of the techniques will be presented in section 4.5.

4.3.3 Stability of Bit-Flipping and Dither

In this section we compare the stability of dithered and DBF modulators using an NTF one-norm technique described by Schreier [Sch92]. The basis of the technique is to derive a relationship between the maximum quantizer error and an upper bound on the quantizer input.

The first stage is to reconfigure the modulator using the noise shaper topology (figure 4.13). The output of the modulator is given by:

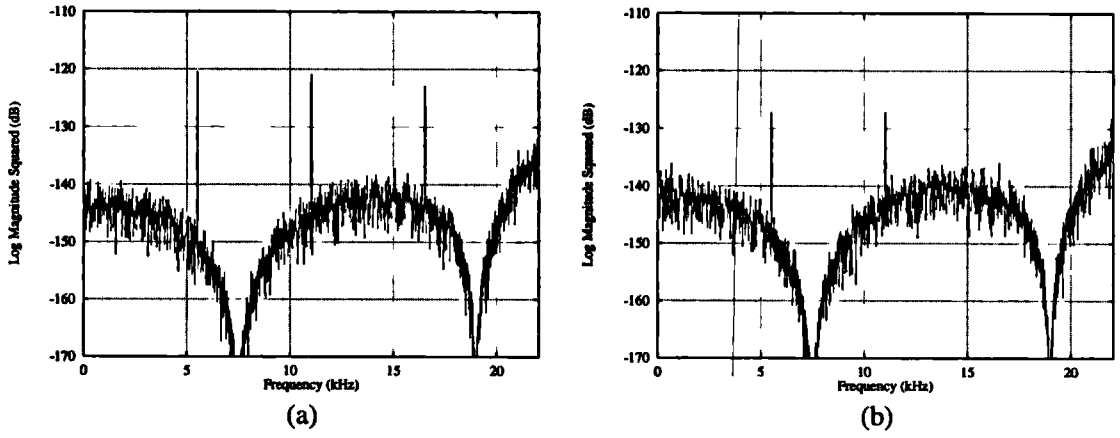


Figure 4.9: Baseband spectrum of DBF modulator with parameters {64, 4, 3.0} and $m_x = 1/512$ for (a) $B = 0$ and (b) $B = 0.04$

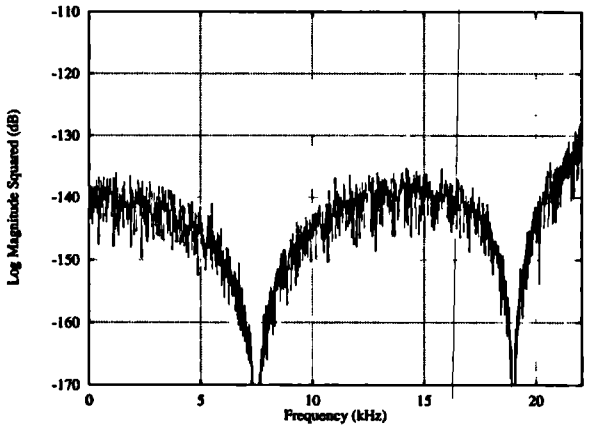


Figure 4.10: Baseband spectrum of DBF modulator with parameters {64, 4, 3.0}, $B = 0.07$ and $m_x = 1/512$

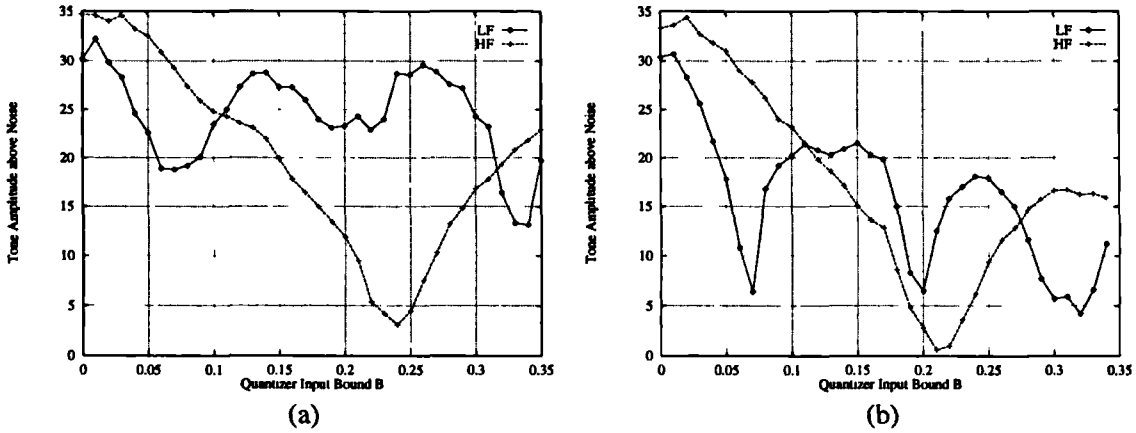


Figure 4.11: Amplitude above noise floor of low and high frequency tones against B for DC input $m_x = 1/512$ and modulator parameters $L = 64$, $P_n = 3.0$ and (a) $N = 3$, (b) $N = 4$

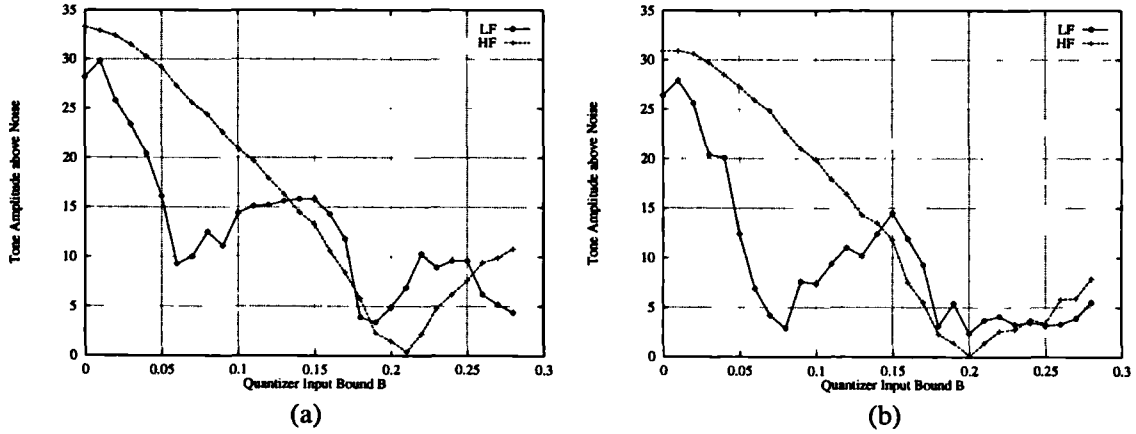


Figure 4.12: Amplitude above noise floor of low and high frequency tones against B for DC input $m_x = 1/512$ and modulator parameters $L = 64$, $P_n = 3.0$ and (a) $N = 5$, (b) $N = 6$

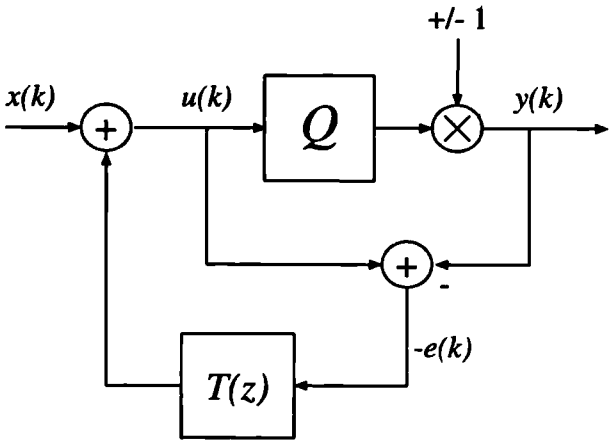


Figure 4.13: Equivalent Noise Shaper Topology of ΣΔ Modulator

$$Y(z) = X(z) + E(z)(1 - T(z)) \tag{4.4}$$

where $E(z) = Y(z) - U(z)$

By setting $T(z) = H(z)/(1+H(z))$, this topology can be made equivalent to the ΣΔ modulator structure, with the exception that the STF is unity.

The time-domain equation for the quantizer input is:

$$u(k) = x(k) + \sum_{i=0}^{\infty} t(i)e(k-i) \tag{4.5}$$

The convolution term is maximised when the signs of $e(k-i)$ match the signs of the filter impulse response $t(i)$, hence $u(k)$ is always bounded as follows:

$$|u(k)| \leq ||x(k)||_\infty + \sum_{i=1}^{\infty} |t(i)||e(k-i)| \tag{4.6}$$

$||x(k)||_\infty$ is the infinity norm of the magnitude of $x(k)$ i.e. the maximum value of its magnitude.

For a quantizer error which is bounded by a constant $|e(k)| \leq \epsilon$, the upper bound for $|u(k)|$ is thus obtained as:

$$||u(k)||_\infty = ||x(k)||_\infty + \epsilon||t(k)||_1 \tag{4.7}$$

$||t(k)||_1$ is the one-norm (i.e. the sum of the magnitudes) of the impulse response of $T(z)$. In [Sch92] a value of $\epsilon = \Delta$ is used as a hypothesised upper bound on the quantizer error, where $\pm\Delta$ are the quantizer levels, and this leads to the derivation

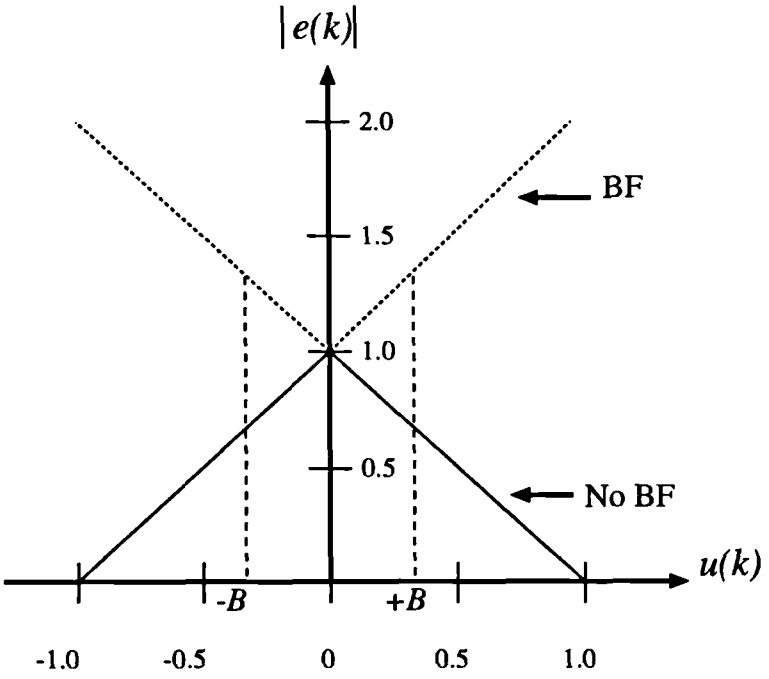


Figure 4.14: Variation in Quantizer Error Magnitude with quantizer input $u(k)$ for bit-flipping and non bit-flipping modulators

of an upper bound for $||t(k)||_1$ for stability, which may be used in the design of the NTF. This bound is very conservative, for example the double loop modulator with $NTF(z) = 1 - z^{-2}$ is deemed unstable by this approach [Sch92].

Equation 4.7 may also simply be used to compare the maximum quantizer input for standard and bit-flipping modulators. For a given $T(z)$, the upper bound of $|u(k)|$ is related by equation 4.7 to the upper bound of the quantizer error ϵ. The quantizer error in each case is plotted against quantizer input in figure 4.14. There are two observations to be drawn from this:

1. With the exception of the case where $u(k) = 0$, the quantizer error for samples in which bit-flipping occur is always greater than without.

2. For samples in which bit-flipping occurs, the quantizer error increases with $u(k)$, whereas without bit-flipping the error reduces with $u(k)$ over the range $-1 < u(k) < 1$, reaching a minimum for $|u(k)| = 1$ where the input falls on the quantizer level.

Due to the increase in quantizer error magnitude, bit-flipping causes ϵ to increase and by equation 4.7, the upper bound on $|u(k)|$ also increases. This can lead to

a further increase in ϵ, triggering modulator instability by positive feedback. As already discussed these bounds are conservative, especially because the BFO does not operate in every sample; however the analysis does indicate that stability is degraded with bit-flipping.

Stability of Deterministic Bit-Flipping and Dither

The above analysis indicates in very general terms that bit flipping can degrade modulator stability by causing a growth in the quantizer error $e(k)$ and quantizer input $u(k)$. Due to the relationship between $||e(k)||_\infty$ and $||u(k)||_\infty$ for bit-flipping modulators, the maximum quantizer error may be reduced by ensuring bit-flipping only occurs for small values of $|u(k)|$. An advantage of the DBF scheme is that bit-flipping will only occur if $|u(k)|$ is bounded by a constant i.e. $|u(k)| < B$. This places an upper bound on $||e(k)||_\infty$ of $1+B$ when bit-flipping occurs i.e. $\epsilon = 1+B$. Therefore the stability of the modulator may be controlled by appropriately choosing of B.

Conceptually, conventional dither also utilises bit-flipping, since the dither signal causes the quantizer state to change occasionally. Using dither, the probability of bit-flipping is greater when $|u(k)|$ is small, and for dither of peak amplitude B (refer to figure 4.7), the maximum quantizer error, $\epsilon = 1 + B$ is the same as with DBF. A crucial difference between the two schemes is that in order to achieve the same bit-flipping rate, the value of B must be larger with dither than DBF, due to the random element in dither which prevents bit-flipping in all samples which satisfy condition 4.3. It will be confirmed in section 4.5, that as a result, dither requires a greater value of B to linearise the modulator.

The probability density function of the dither influences the value of B required for a given bit-flipping rate. It will be shown below that one-bit dither has a 50% probability of bit-flipping if condition 4.1 is satisfied. Rectangular has a smaller probability as the distribution is spread between $+B$ and $-B$. Triangular PDF dither has an even smaller probability of bit-flipping because it has a smaller probability of high values of $r(k)$. For the same bit-flipping rate, the required value of B and therefore the value of ϵ increases in the order DBF < 1-bit dither < rectangular PDF < triangular PDF. This ties up with experimental results in section 4.5 which determine the relative stabilities of the different linearisation schemes.

4.3.4 Relationship between One-bit Dither and Bit-Flipping

It has been shown in section 4.2 that a dither signal added to the quantizer input causes an equivalent bit-flipping operation on the quantizer output. This concept is now used to derive a relationship between single bit dither and DBF. The single-bit dither input is defined as follows

$$r(k) = \begin{cases} +B & \text{if } R \geq 0.5 \\ -B & \text{if } R < 0.5 \end{cases} \tag{4.8}$$

where R is a uniformly distributed random variable: $0 < R < 1$

There are two cases from conditions 4.1 and 4.2:

- Case 1: $|u(k)| \geq B$
  There is no possibility of the dither flipping the quantizer output state.

- Case 2: $|u(k)| < B$
  The dither will flip the output state if either: $0 < u(k) < B$ and $r(k) = -B$
  or: $-B < u(k) < 0$ and $r(k) = +B$

Since $p\{r(k) = -B\} = p\{r(k) = +B\} = 0.5$, we obtain the bit-flipping condition:

$$\textit{Flip output if } |u(k)| < B \textit{ and } R < 0.5$$

Therefore single bit dither is equivalent to DBF with the inclusion of a random condition 'in-line' with the BFO. This is clarified by the implementation details described in appendix C.2.

4.4 Quasi-linear Analysis of Dithered and Deterministic Bit-Flipping Modulators

The aim of this section is to model the behaviour of $\Sigma\Delta$ modulators employing dither or deterministic bit-flipping in order to compare the noise performance on a theoretical basis. The technique of quasi-linear analysis using the PDF method, introduced in chapter 3 is used as the basis for the modelling. Quasi-linear (i.e.

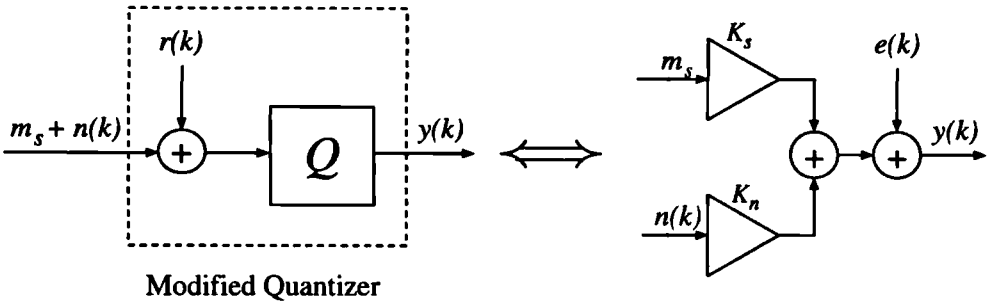


Figure 4.15: Quasi-linear model of dithered quantizer

statistical) analysis is especially suitable for dithered modulators since there is the inclusion of a random input and this tends to improve the accuracy of the Gaussian assumption. The quasi-linear technique has been used previously in the modelling of $\Sigma\Delta$ modulation with a noise input [Ard88] and [Sto90] and a noise+DC input [Ker96].

The fundamental approach taken here is that the dither and DBF is modelled as a modification to the quantizer nonlinearity. This approach is based upon proposition 4.1: that both techniques modify the decision made by the quantizer for some samples. It will be shown that the bit-flipping modifies the statistics of the nonlinearity, causing the gain and error variance to change. This modelling approach differs notably from previous analysis, for example [Ard88], where an additional noise source at the system input is modelled using an additional gain term i.e. two separate noise gains are used.

With the exception of the modified nonlinearity, the conceptual model for dither and DBF is identical to the undithered modulator. For both the dither and DBF model, the quantizer input signal is separated into AC and DC components, which are passed through separate gain terms before being added to form an estimate of the quantizer output (figures 4.15 and 4.16). The error introduced by the modified quantizer is defined as the difference between the estimate of the modified quantizer output and the actual modified quantizer output, and this error is included in the model as an additive noise source (refer to chapter 3).

In the following, a set of nonlinear equations are derived which describe the behaviour of the two types of modulator. Only an outline derivation is given, though further details are available in appendix C.3. As with the previous quasi-linear analysis the assumption is made that the signals are ergodic and quasi-stationary.

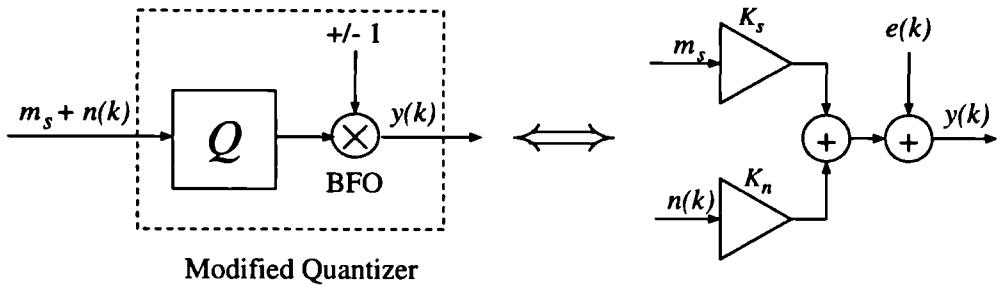


Figure 4.16: Quasi-linear model of DBF quantizer

4.4.1 Dithered Modulation

The analysis here is restricted to the case of rectangular PDF dither, as this is sufficient to demonstrate the principles and allow comparisons; the analysis is general enough to be extended readily to other dither types such as one-bit and triangular dither if required.

We begin with a definition of the quantizer nonlinearity:

$$\mathcal{Q}(n(k)+r(k)+m_s) = \begin{cases} \Delta & n(k)+r(k)+m_s \geq 0 \\ -\Delta & n(k)+r(k)+m_s < 0 \end{cases} \tag{4.9}$$

The assumption is made that the AC component of the quantizer input, $n(k)$, has a Gaussian PDF, and of course the dither source $r(k)$ has a rectangular PDF.

$$p(n) = \frac{1}{\sigma_n\sqrt{2\pi}} e^{-n^2/2\sigma_n^2} \tag{4.10}$$

$$p(r) = \begin{cases} \frac{1}{2\delta} & -\delta < r < \delta \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

where δ is the peak dither amplitude (equal to B in figure 4.7).[1]

The quantizer error signal is given by:

$$e(k) = \mathcal{Q}\{n(k)+r(k)+m_s\} - n(k)K_n - m_sK_s \tag{4.12}$$

and its variance σ_e^2 is obtained from the joint statistics of the dither and input signal.

$$\sigma_e^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{Q}\{n+r+m_s\} - nK_n - m_sK_s\}^2 p(n)p(r)dndr \tag{4.13}$$

[1]This dither notation is used to avoid confusion in subsequent comparisons.

K_n and K_s are chosen to minimise the variance σ_e^2 by partial differentiation, giving:

$$K_n = \frac{1}{\sigma_n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{Q}\{n+r+m_s\}np(n)p(r)dndr \tag{4.14}$$

$$K_s = \frac{1}{m_s}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{Q}\{n+r+m_s\}p(n)p(r)dndr \tag{4.15}$$

Referring to appendix C.3, it can be shown that:

$$K_n = \frac{\Delta}{2\delta}\left[erf\left(\frac{\delta+m_s}{\sigma_n\sqrt{2}}\right)+erf\left(\frac{\delta-m_s}{\sigma_n\sqrt{2}}\right)\right] \tag{4.16}$$

$$K_s = \frac{\Delta}{2\delta m_s}\left\{(\delta+m_s)erf\left(\frac{\delta+m_s}{\sigma_n\sqrt{2}}\right)+(\delta-m_s)erf\left(\frac{\delta-m_s}{\sigma_n\sqrt{2}}\right)\right\} + \frac{\Delta\sigma_n}{\delta m_s\sqrt{2\pi}}\left\{e^{-\frac{(\delta+m_s)^2}{2\sigma_n^2}}-e^{-\frac{(\delta-m_s)^2}{2\sigma_n^2}}\right\} \tag{4.17}$$

We define:

$$\alpha = \frac{\delta+m_s}{\sigma_n\sqrt{2}} \tag{4.18}$$

$$\beta = \frac{\delta-m_s}{\sigma_n\sqrt{2}} \tag{4.19}$$

Since $m_y = m_s K_s$, equation 4.17 can be rewritten

$$m_y = \frac{\Delta}{2\delta}\{(\delta+m_s)erf(\alpha)+(\delta-m_s)erf(\beta)\} + \frac{\Delta\sigma_n}{\delta\sqrt{2\pi}}\left\{e^{-\alpha^2}-e^{-\beta^2}\right\} \tag{4.20}$$

The quantizer error variance σ_e^2 is obtained by substituting the expression for K_n into equation 3.19:

$$\sigma_e^2 = \Delta^2 - K_n^2\sigma_n^2 - m_y^2 = \Delta^2 - \frac{\Delta^2\sigma_n^2}{4\delta^2}[erf(\alpha)+erf(\beta)]^2 - m_y^2 \tag{4.21}$$

Note that unlike the undithered modulator it is not possible to obtain a closed-form equation for σ_e^2 which is independent of σ_n^2; however σ_n^2 and m_s^2 are linked by equation 4.20 and are therefore independent of K_n. This means that the white noise assumption is not used in determining σ_e^2.

To evaluate K_n the white noise assumption must be used to link σ_n^2 and σ_e^2:

$$\sigma_n^2 = \sigma_e^2 P_s(K_n) \tag{4.22}$$

where $P_s(K_n)$ is defined in equation 3.27.

Substituting equation 4.16 and 4.21 into 4.22 we obtain:

$$\left[\Delta^2 - \frac{\Delta^2\sigma_n^2}{4\delta^2}\left[erf(\alpha) + erf(\beta)\right]^2 - m_y^2\right] P_s\left(\frac{\Delta}{2\delta}\left[erf(\alpha) + erf(\beta)\right]\right) - \sigma_n^2 = 0 \tag{4.23}$$

From equation 4.20 we obtain:

$$\frac{\Delta}{2\delta}\left\{(\delta + m_s)erf\,(\alpha) + (\delta - m_s)erf\,(\beta)\right\} + \frac{\Delta\sigma_n}{\delta\sqrt{2\pi}}\left[e^{-\alpha^2} - e^{-\beta^2}\right] - m_y = 0 \tag{4.24}$$

The equations are now in a form which can be solved using a nonlinear equation solver (refer to appendix B.2) to obtain values of m_s and σ_n^2 for the fixed parameters m_y and δ:

$$\begin{aligned} f_0(m_s, \sigma_n) &= 0 \qquad \text{(equation 4.23)} \\ f_1(m_s, \sigma_n) &= 0 \qquad \text{(equation 4.24)} \end{aligned} \tag{4.25}$$

These values may then be back-substituted into equations 4.16 and 4.21 to obtain values for K_n and σ_e^2

The quasi-linear parameters can alternatively be calculated by time-averages in exactly the same manner as in section 3.3.2, noting that $u(k)$ is now the input to the dithered quantizer, i.e. the signal *before* the dither source.

4.4.2 DBF Modulator

The analysis of the DBF modulator is more straightforward because there is no random noise source and the integral expressions are simpler.

The only difference between the undithered and DBF modulator is the quantizer nonlinearity:

$$\mathcal{Q}\{n(k)+m_s\} = \begin{cases} \Delta & n(k)+m_s \geq B \\ -\Delta & 0 \leq n(k)+m_s < B \\ \Delta & -B \leq n(k)+m_s < 0 \\ -\Delta & n(k)+m_s < -B \end{cases} \tag{4.26}$$

The signal and noise gains, are obtained by substituting the new quantizer non-linearity into equations 3.12 and 3.11, assuming the AC component of the quantizer input has a Gaussian distribution (refer to appendix C.3 for a full derivation):

$$K_n = \frac{2\Delta}{\sigma_n\sqrt{2\pi}}\left[e^{-\frac{(B+m_s)^2}{2\sigma_n^2}} + e^{-\frac{(B-m_s)^2}{2\sigma_n^2}} - e^{-\frac{m_s^2}{2\sigma_n^2}}\right] \tag{4.27}$$

As a consistency check, putting $B = 0$ leads to the cancellation of the last two exponential terms and the same result as an undithered modulator is obtained (equation 3.15).

The signal gain is given by:

$$K_s = \frac{\Delta}{m_s}\left[erf\left(\frac{B+m_s}{\sigma_n\sqrt{2}}\right) - erf\left(\frac{B-m_s}{\sigma_n\sqrt{2}}\right) - erf\left(\frac{m_s}{\sigma_n\sqrt{2}}\right)\right] \tag{4.28}$$

Again, as a consistency check, putting $B = 0$ leads to the cancellation of the first and last $erf()$ terms and, noting that $erf()$ is an odd function, the same result as an undithered modulator is produced (equation 3.16).

We define:

$$\gamma = \frac{B+m_s}{\sigma_n\sqrt{2}} \tag{4.29}$$

$$\lambda = \frac{B-m_s}{\sigma_n\sqrt{2}} \tag{4.30}$$

$$\rho = \frac{m_s}{\sigma_n\sqrt{2}} \tag{4.31}$$

$$\tag{4.32}$$

Noting that $m_y = K_s m_s$, equation 4.28 can be rewritten as:

$$m_y = \Delta\left[erf\left(\gamma\right) - erf\left(\lambda\right) - erf\left(\rho\right)\right] \tag{4.33}$$

The quantizer error variance σ_e^2 is obtained by substituting the expression for K_n into equation 3.19:

$$\begin{aligned}\sigma_e^2 &= \Delta^2 - K_n^2\sigma_n^2 - m_y^2 \\ &= \Delta^2 - \frac{2\Delta^2}{\pi}\left[e^{-\gamma^2} + e^{-\lambda^2} - e^{-\rho^2}\right]^2 - m_y^2 \end{aligned} \tag{4.34}$$

Again it is not possible to obtain a closed-form equation for σ_e^2 which is independent of σ_n^2; however σ_n^2 and m_s^2 are linked by equation 4.33 and are therefore independent of K_n. Therefore the white noise assumption is not used in determining σ_e^2.

As before, using the white noise approximation:

$$\sigma_n^2 = \sigma_e^2 P_s(K_n) \tag{4.35}$$

Equation 4.27 and 4.34 are substituted into 4.35 to obtain:

$$\left[\Delta^2 - \frac{2\Delta^2}{\pi}\left[e^{-\gamma^2} + e^{-\lambda^2} - e^{-\rho^2}\right] - m_y^2\right] P_s\left(\frac{2\Delta}{\sigma_n\sqrt{2\pi}}\left[e^{-\gamma^2} + e^{-\lambda^2} - e^{-\rho^2}\right]\right) - \sigma_n^2 = 0 \tag{4.36}$$

$P_s(.)$ is defined in equation 3.27.

From equation 4.33 we obtain:

$$\Delta\left[erf\left(\gamma\right) - erf\left(\lambda\right) - erf\left(\rho\right)\right] - m_y = 0 \tag{4.37}$$

The equations are in a form which can be solved using a nonlinear equation solver (refer to appendix B.2) to obtain values of m_s and σ_n^2 for the fixed parameters m_y and δ:

$$\begin{aligned} f_0(m_s, \sigma_n) &= 0 \qquad \text{(equation 4.36)} \\ f_1(m_s, \sigma_n) &= 0 \qquad \text{(equation 4.37)} \end{aligned} \tag{4.38}$$

These values are substituted back into equations 4.27 and 4.34 to obtain values for K_n and σ_e^2.

As with the undithered and dithered modulators, the quasi-linear parameters can alternatively be calculated by time-averages.

For both dithered and DBF modulators, the values of K_n and σ_e^2 obtained by the solution of the nonlinear equations can be used to obtain an estimate of the baseband noise power by evaluating equation 3.31:

$$P_b = \frac{\sigma_e^2}{\pi} \int_0^{\pi/L} \left| \frac{1}{1 + H(\theta)K_n} \right|^2 d\theta \tag{4.39}$$

4.4.3 Examples of Quasi-linear Analysis of Dithered and DBF Modulators

In the following examples the quasi-linear model with dither and DBF has been tested by solving the system of equations 4.25 and 4.38 for different values of δ and B for a DC input level $m_x = 0.1$, and modulator parameters $\{64, 4, 3.0\}$. The quasi-linear parameters obtained using the PDF method are compared to the time-average method.

In figure 4.17(a) and (b) the quasi-linear gain K_n is plotted against the δ and B for dithered and DBF modulators, respectively. The value of K_n reduces with δ and B. Again two solutions to K_n exist which correspond to the stable and unstable equilibrium in the $P_n(K_n)$ curve (refer to section 3.5 of chapter 3). In these results, only the solution corresponding to a stable equilibrium is shown. On the same axis the simulated values of K_n obtained using 100,000 time samples are plotted. For both modulator types, K_n falls with δ and B. In figure 4.18 σ_e^2 is plotted against δ and B for the dithered and DBF modulator. As the parameter δ and B increases, σ_e^2 increases.

This behaviour is expected - an increase in either δ or B causes the the bit-flipping activity to increase. This causes the maximum value of the quantizer error to increase, and the maximum value of the quantizer input to increase (refer to equation 4.7 in section 4.3.3).

As the bit-flipping rate increases and K_n reduces, the minimum value of the $P_n(K_n)$ curve is reached. Beyond this there is no solution to the system of equations and this represents modulator instability (refer to chapter 3). For dither, the DC MSA predicted by the PDF method is considerably larger than observed in practice; however in the case of DBF the MSAs are similar. This requires further investigation, however a possible reason is that the predictions of the quasi-linear model are based upon average signal characteristics, whereas in practice the peak signal level of the dither may drive the modulator into instability. The large peak dither amplitude causes a large peak quantizer error which may may cause the modulator to become unstable due to attraction to regions in state-space corresponding to saturation limit

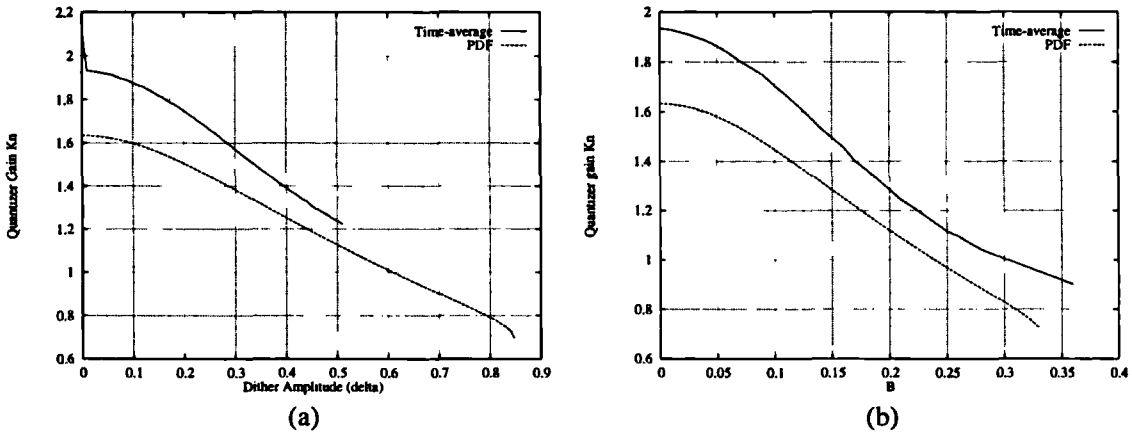


Figure 4.17: (a) K_n vs δ and (b) K_n vs B for dithered and DBF modulators with parameters {64, 4, 3.0} and $m_x = 0.1$

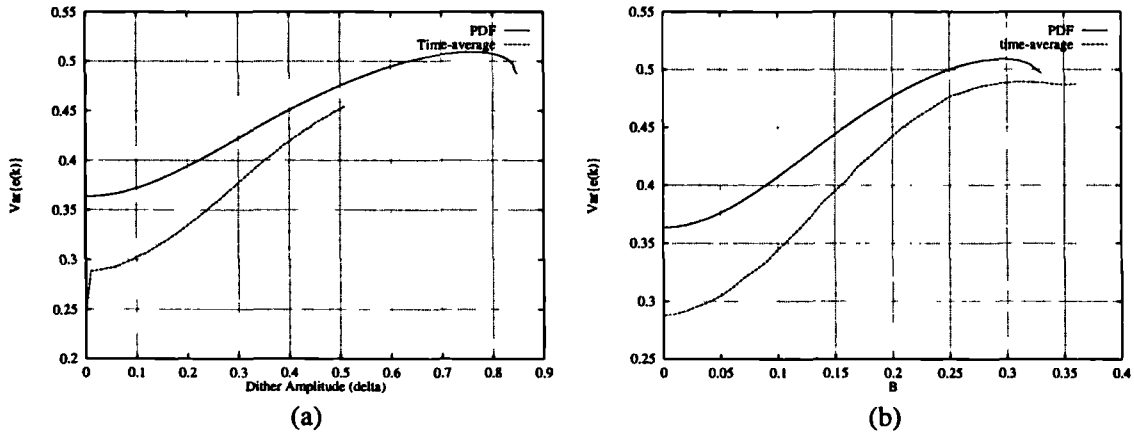


Figure 4.18: (a) σ_e^2 vs δ and (b) σ_e^2 vs B for dithered and DBF modulators with parameters {64, 4, 3.0} and $m_x = 0.1$

cycles. This is not as likely to happen with DBF, where the maximum value of the quantizer error is bounded (refer to section 4.3.3).

In figure 4.19, the baseband noise power is plotted against the parameter δ and B. Due to the increase in σ_e^2 and decrease in K_n, the baseband noise power increases with B for both dither and DBF. Also plotted along with these curves is the noise power obtained by the spectral estimation method (refer to section A.1), which provides a reference curve.

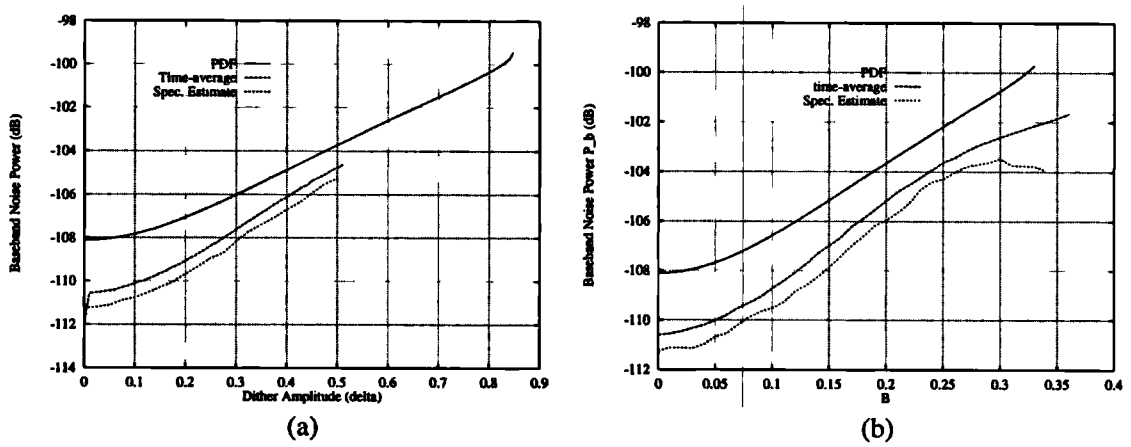


Figure 4.19: Baseband Noise Power against (a) δ and (b) B for dithered and DBF modulator with parameters {64, 4, 3.0} and $m_x = 0.1$

Errors in Gaussian and White Noise Assumptions

Notice that for the dithered modulator, the errors in the time-average and PDF curves reduces as δ increases, as observed in the curves for K_n, σ_e^2 and P_b. This implies that the Gaussian PDF assumption becomes more accurate. As evidence of this, the PDFs of the quantizer input are shown in figure 4.20 for dither levels $\delta = 0$ and $\delta = 0.5$. Plotted on the same axis are the theoretical distributions of a Gaussian AC input with DC offset of m_s and variance σ_n^2, as measured in the simulations by the time-average method (strictly this is the Normal distribution with mean m_s). These results show that adding dither causes the AC component of the quantizer input to become more Gaussian. Note that the PDF measurement is of the *modified* quantizer input, i.e. before the dither source.

For the DBF modulator the errors between the Gaussian and time-average curves for σ_e^2 reduce as the value of B increases, implying that the Gaussian assumption becomes more accurate. A different characteristic is observed for K_n, with the errors initially reducing with B then increasing at greater input levels. The divergence at high input levels is possibly due to the white noise assumption becoming less accurate (note that the white noise assumption is not used to determine σ_e^2). As further evidence of this, it can be seen in figure 4.19(b) that the error between the time-average curve, which uses only the white noise assumption, and the spectral estimation curve increases for high levels of B.

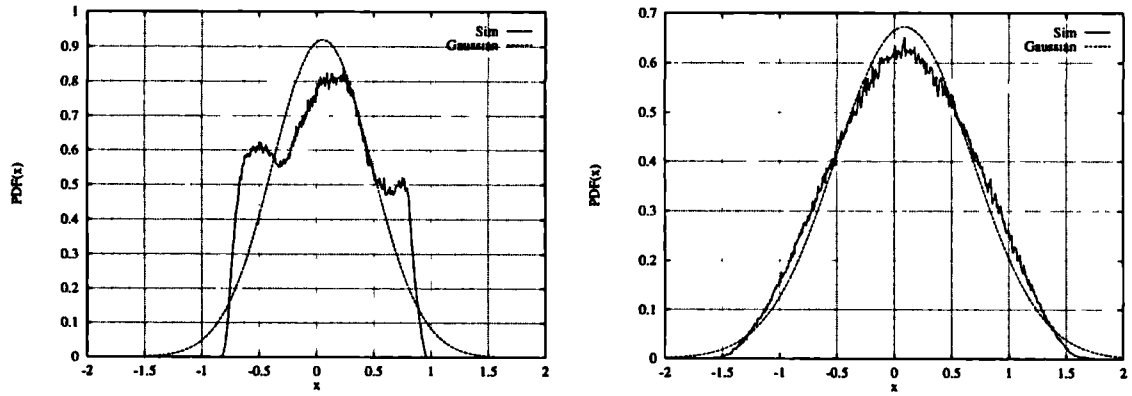


Figure 4.20: PDF of quantizer input for dithered modulator with $m_y = 0.1$ and {64, 4, 3.0} for (a) $\delta = 0$ and (b) $\delta = 0.5$

4.4.4 Dither and DBF Duality

The similarity between the effect of dither and DBF on the quasi-linear parameters and the baseband noise power is now investigated more closely. In figures 4.21 and 4.22 the values of K_n, σ_e^2 and the baseband noise power P_b obtained using the PDF method are plotted against the parameter s, where $s = 100\delta = 256B$, for both dither and DBF. The modulator parameters are {64, 4, 3.0}. These results show that for the tested modulator, making the Gaussian and white noise assumptions, for $\delta = 2.56B$, dither and DBF have an almost identical effect on the behaviour of the modulator, resulting in an almost identical predicted noise power. Furthermore the predicted stability of the two schemes is almost identical. This implies that in terms of dynamic range, the theoretical performance of the two systems is almost identical. In practice, however, the stability of the dithered modulator is considerably worse than predicted, for the reasons described in section 4.4.3.

4.5 Investigations and Results

In this section a comparison of the following linearisation schemes are presented:

- Type 1: Rectangular PDF dither of peak amplitude B.
- Type 2: Triangular PDF high-pass dither of peak amplitude B, obtained by prefiltering rectangular PDF dither of peak amplitude $B/2$ with transfer function $1 - z^{-1}$.

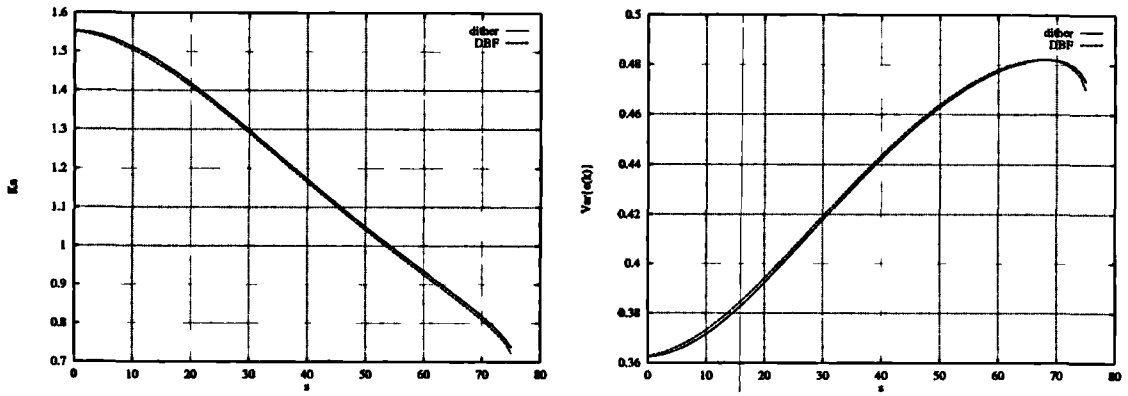


Figure 4.21: Comparison of (a) K_n and (b) σ_e^2 for dithered and DBF modulators $\{64, 4, 3.0\}$ with $m_y = 0.1$ and $\delta = 2.56B$

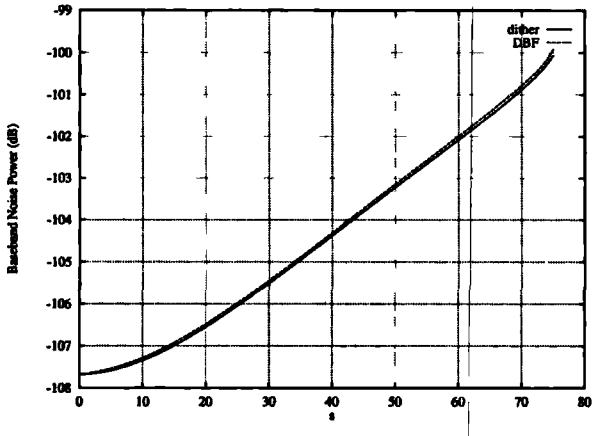


Figure 4.22: Comparison of Baseband Noise Power P_b for dithered and DBF modulator $\{64, 4, 3.0\}$ with $m_y = 0.1$ and $\delta = 2.56B$

- Type 3: One-bit dither with amplitude $\pm B$.

- Type 4: Deterministic Bit-Flipping with quantizer input bound B.

The aim is to establish the success at which each scheme linearises the modulator and in particular, to establish the relative dynamic range penalties. Since the linearity of the modulator is dependent on its order, results are presented for modulators of order three to six. The investigation is restricted to orders above two, since these are more suitable for audio conversion due to their wide dynamic range at moderate oversampling ratios. To simplify the comparison, the study is also restricted to the case of oversampling ratio $L = 64$. As this oversampling ratio is very common for audio converters, the results presented have practical value.

4.5.1 Maximum Power Gain NTFs

The quasi-linear analysis in section 4.4 has shown that the application of dither or DBF influences both the noise power and stability of the modulator. Although it is simple to calculate the dynamic range penalty of each scheme using modulators with the same power gains, a comparison on this basis is not strictly adequate. Referring to section 2.3.1 a tradeoff exists between the MSA and noise power, and empirical results show that the dynamic range tends to reach a maximum for modulators with a MSA in the range 0.25-0.35 (this has also been pointed out by Risbo [Ris94b]). Therefore the dynamic range is strongly dependent on the MSA and comparing modulators with vastly differing MSAs is not appropriate. Furthermore, from a design perspective it is more useful to compare the performance of modulators which have the same MSA, which is chosen in advance to meet a particular specification.

The results here are based upon modulators with a MSA of 0.3 which is within the optimal range for maximum dynamic range. The design procedure of appendix A.4 is used to obtain the NTF with the maximum power gain for stability. The maximum power gain is P_m. This parameter is also a measure of the stability of the modulator with the given bit-flipping scheme i.e. a higher P_m means the modulator has inherently higher stability margins. It should be emphasised that a modulator which has a lower P_m will naturally have increased baseband noise. In other words, in deriving modulators with the same maximum input level, we are trading off stability against noise performance.

DC input	Measured Tone Frequencies			
	LF			HF
1/512	5512.5	11025.0	16537.5	1408443.75
1/256	11025.5	22051.0	-	1405687.5

Table 4.1: Measured tone Frequencies for different DC inputs

4.5.2 Linearity Evaluation

In section 2.4 it was explained that the two main artifacts of nonlinear behaviour are: tonal behaviour with DC inputs, and noise modulation. Either of these can be used to assess modulator linearity. Assessing the performance of a given linearisation scheme requires a large number of simulations to establish an appropriate level of B for successful linearisation. As a consequence, it is important to assess the linearity using the most efficient method. Determining levels of noise modulation is extremely inefficient — for a detailed evaluation, it has been found that simulations covering a dynamic range of 120 dB in steps of 0.5 dB are required. Each simulation requires 1.5 million time steps to ensure that tonal behaviour is reliably identified. This means that approximately 360 million time steps are required for each dither level tested.

A much quicker method that has been found to achieve reliable results is to perform a spot check of the tonal performance for two rational DC input levels which cause tones in the baseband region. For these tests, the two DC inputs used are $m_1 = 1/512$ and $m_2 = 1/256$. Assuming that the phase-inversion pattern dominates (refer to section 2.4.1), for an oversampling ratio $L = 64$, the fundamental baseband tones for the two inputs will fall at frequencies of 5512.5 kHz and 11025 kHz respectively. Additionally, harmonics may exist which fall into the baseband. For each input, the tones at the frequencies given in table 4.1 have been measured.

In section 2.4.2 it was explained that high frequency tones can be problematic due to intermodulation with spurious clock frequencies. Therefore the amplitude of the high frequency tone is also measured, for a DC input $m_x = 1/512$. The tone frequency is given in table 4.1.

Linearity Criterion

The most meaningful way of measuring the tone amplitudes is to make the measurement relative to the noise floor. This allows a comparison of the linearity of modulators with different noise-shaping characteristics and orders, which have different levels of baseband noise attenuation. The measurement technique is described in appendix A.3.

To compare the different schemes it is necessary to define a reference linearity which each scheme must achieve. For this study, the reference linearity is a maximum tone amplitude of 4 dB relative to the noise floor. Modulators with any tone more than 4 dB above the noise floor are considered to be 'tonal' and require more linearisation. This level is fairly arbitrary, though it allows a comparison between the different schemes, as long as the FFT size is constant for the comparisons (the FFT parameters are given in appendix A.1). Note the 4 dB criteria is used also in [Mot96] as an audibility criterion, though no reference is made to the FFT size.

For each linearisation scheme and order, the following steps used in obtaining a modulator with a reference linearity and MSA. The procedure is performed twice to obtain modulators optimised for baseband and wideband linearity. For the baseband linearity evaluation, only the tones which fall into the baseband are used to evaluate the linearity criterion. For wideband linearity, both the baseband tones and the high tone are evaluated.

1. Begin with $B = 0$.
2. Design the maximal power gain NTF for an MSA of 0.3.
3. Evaluate the modulator linearity.
4. If sufficient linearity is not achieved. increase B in steps of 0.01 and repeat from step 2.

Once optimal dither parameters have been found, a single noise modulation test is performed to double-check the linearity.

4.5.3 Summary of Results

In this section, a summary of results is presented. For each of the four schemes, the minimum dither amplitude required to obtain a non-tonal modulator (as defined

Order	P_m	Max LF Tone	Max HF Tone	Baseband Noise Power (dB)
3	4.45	27.2	25.2	-102.0
4	4.25	23.1	24.1	-119.9
5	4.10	21.1	24.0	-132.9
6	4.00	15.4	23.7	-144.7

Table 4.2: Performance of undithered modulator

	Baseband				Wideband			
Order	P_m	B	Maximum LF Tone	DR Pen. (dB)	P_m	B	HF Tone	DR Pen. (dB)
3	3.95	0.27	3.1	2.3	2.9	0.7	3.8	13.2
4	3.70	0.21	3.0	5.1	2.70	0.58	2.9	18.6
5	3.75	0.17	2.4	4.0	2.55	0.53	2.9	22.8
6	3.70	0.16	4.0	4.7	2.70	0.48	3.7	22.9

Table 4.3: Performance of Rectangular PDF dither

above) has been obtained, under the maximum input level constraints described in appendix A.4. Note that for the DBF scheme, the results for baseband attenuation correspond to the region around the first minimum in the characteristic curve of figures 4.11 and 4.12, and the results for wideband attenuation correspond to the region around the second minimum. The results are presented in tables 4.2 to 4.6. In the tables, all the tone amplitudes are given in dB. The dash (-) indicates that the scheme failed the linearity criterion. The key comparative measure is the dynamic range (DR) penalty, which indicates the reduction in dynamic range over an undithered maximum power gain modulator. For this measurement the baseband noise power is measured for the DC input level of 1/512. Another important measurement is the maximum NTF power gain P_m, which provides a measure of the relative stabilities of the different schemes. Comparing two dither schemes, the one with the higher P_m has larger stability margins, since it can accept a greater NTF noise amplification.

The results presented in the tables have been confirmed by visual examination of the baseband and wideband FFTs. A sample of the wideband, baseband and noise modulation results are presented in figures 4.23 to 4.29. These results are for the fifth order modulator using parameters taken from the tables. In these plots the

	Baseband				Wideband			
Order	P_m	B	Maximum LF Tone	DR Pen. (dB)	P_m	B	HF Tone	DR Pen. (dB)
3	3.85	0.42	2.5	3.2	2.55	1.24	3.9	16.2
4	3.6	0.32	2.5	5.9	2.35	1.00	3.6	22.8
5	3.5	0.28	2.4	6.8	2.30	0.88	3.4	26.7
6	3.6	0.2	2.3	5.5	2.30	0.82	3.1	30.6

Table 4.4: Performance of Triangular PDF dither

	Baseband				Wideband			
Order	P_m	B	Maximum LF Tone	DR Pen. (dB)	P_m	B	HF Tone	DR Pen. (dB)
3	4.05	0.14	3.9	1.1	-	-	-	-
4	3.65	0.13	3.1	5.6	2.85	0.32	3.6	16.3
5	3.65	0.13	3.3	5.7	2.80	0.30	2.1	18.9
6	3.65	0.10	4.0	5.4	2.70	0.30	2.7	23.0

Table 4.5: Performance of 1-bit dither

	Baseband				Wideband			
Order	P_m	B	Maximum LF Tone	DR Pen. (dB)	P_m	B	HF Tone	DR Pen. (dB)
3	-	-	-	-	-	-	-	-
4	3.75	0.08	3.4	4.6	3.1	0.21	0.2	14.7
5	3.7	0.07	3.4	4.6	3.05	0.2	2.0	16.4
6	3.7	0.07	3.7	5.0	2.95	0.18	3.6	18.8

Table 4.6: Performance of DBF

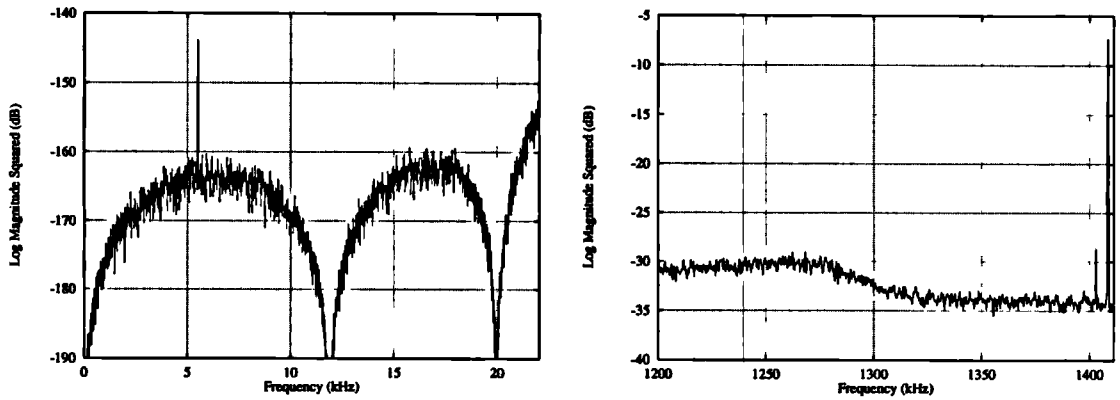


Figure 4.23: (a) Baseband Spectrum of undithered fifth order modulator, $m_x = 1/512$. (b) Zoomed Wideband Spectrum

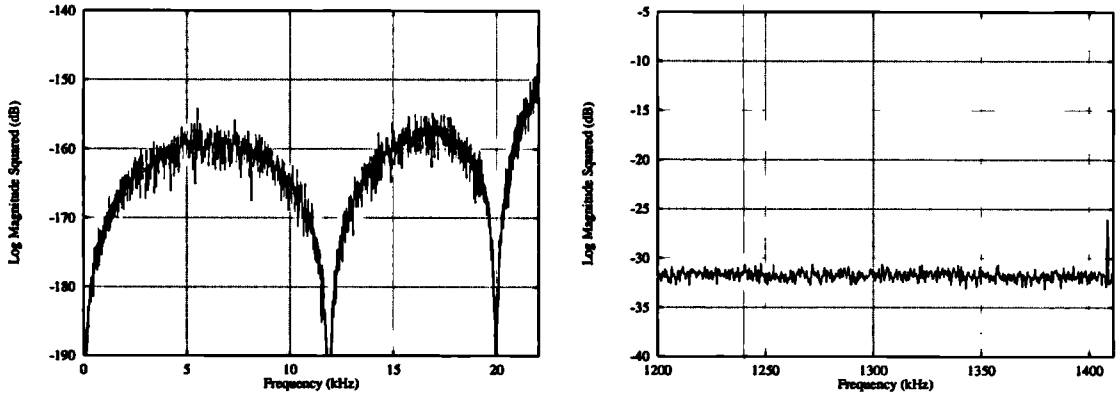


Figure 4.24: (a) Baseband Spectrum and of Fifth order modulator linearised with Rectangular PDF dither, $m_x = 1/512$. (b) Zoomed Wideband Spectrum

baseband and wideband spectra are for different modulators which are optimised for baseband and wideband linearity, respectively. From the tabulated results the following general observations have been made:

- The dither level required to attenuate the high frequency (HF) tone is higher than for the baseband tone. Consequently, for the HF tone P_m is lower and the noise penalty is higher.
- The dither levels generally reduce with order, however the DR penalty increases with order. The increase is sharper for the high frequency tone.
- The tone amplitudes of the linearised modulators are not identical (i.e. not

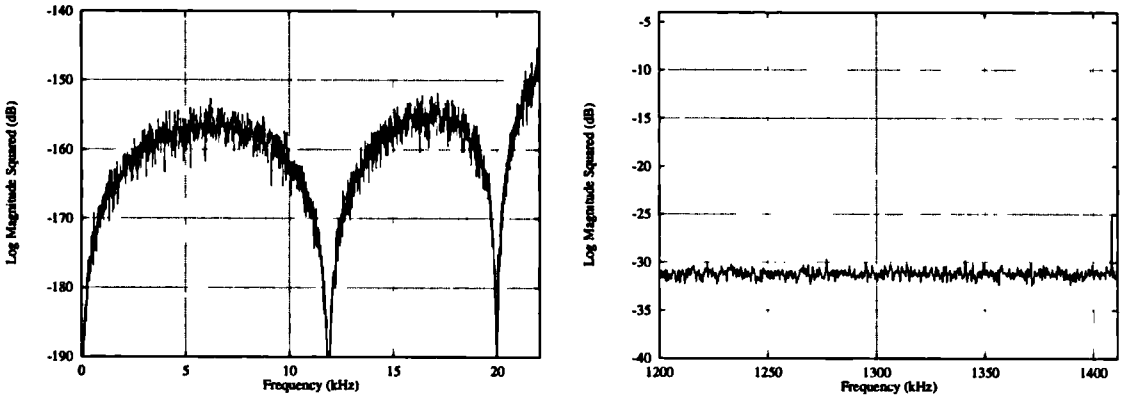


Figure 4.25: (a) Baseband Spectrum of Fifth order modulator linearised with Triangular PDF dither, $m_x = 1/512$. (b) Zoomed Wideband Spectrum.

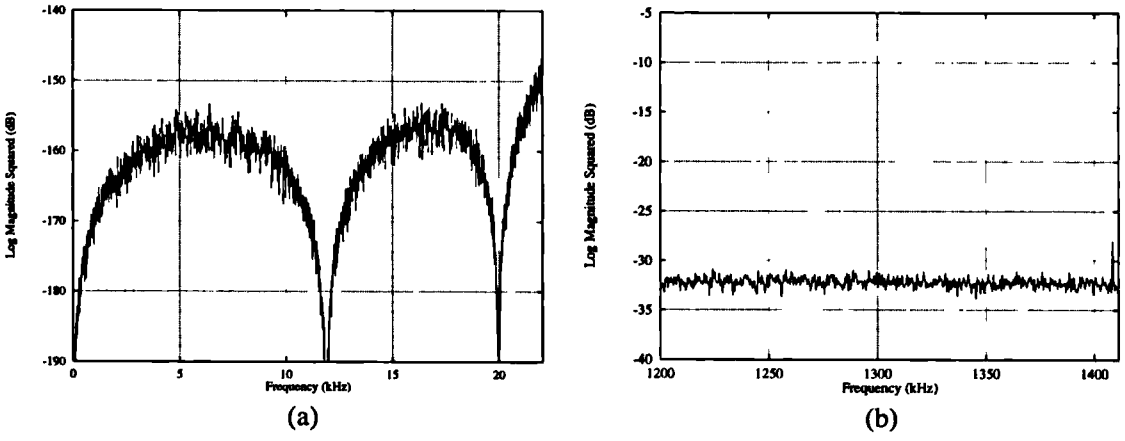


Figure 4.26: (a) Baseband Spectrum of Fifth order modulator linearised with one-bit dither, (b) Zoomed Wideband Spectrum

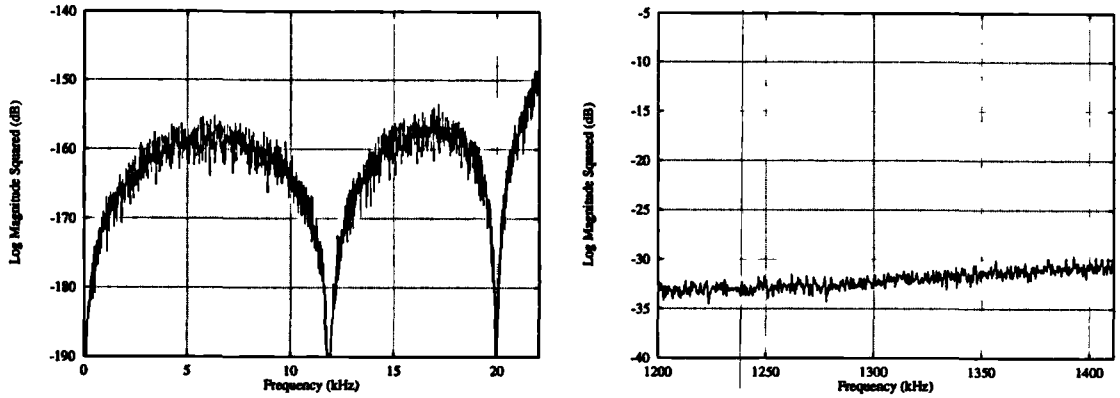


Figure 4.27: (a) Baseband Spectrum of Fifth order modulator linearised with DBF, (b) Zoomed Wideband Spectrum.

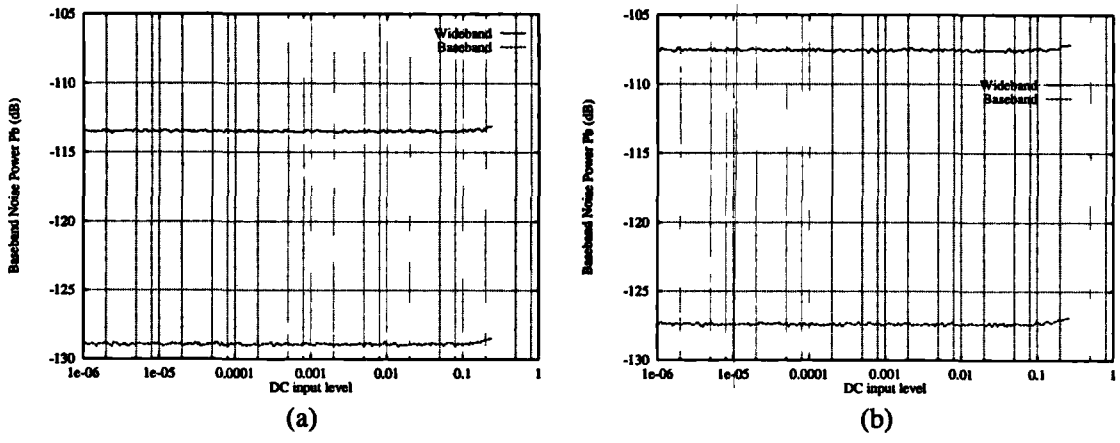


Figure 4.28: Baseband and Wideband Noise modulation plots of 5th order modulator with (a) rectangular PDF dither (b) Triangular PDF dither.

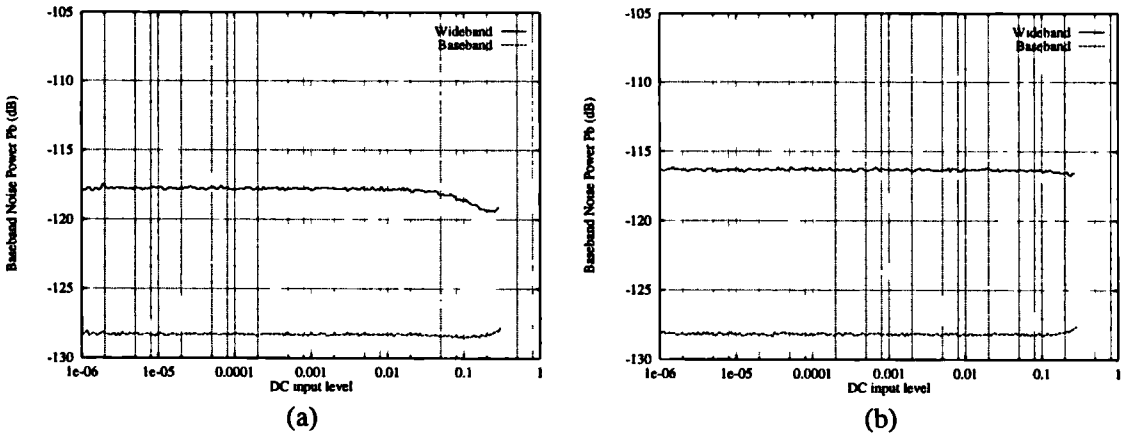


Figure 4.29: Baseband and Wideband Noise modulation plots of 5th order modulator with (a) DBF (b) One-bit dither.

all 4 dB) due to the quantization of the dither level to two decimal places, therefore there will be a slight error in the relative penalties.

- For both baseband and wideband linearisation, the maximum power gain P_m increases in the order Triangular PDF < Rectangular PDF < 1-Bit < DBF. This indicates that the relative stabilities all increase in this order. This result corresponds inversely to the increase in the quantizer error bound ϵ, described in section 4.3.3.

Note that it is not possible to directly compare stability defined by P_m *across* modulator orders without converting the power gains to the values of $min\{P_n(K_n)\}$ obtained by quasi-linear analysis (refer to section 3.5.1).

More concise comparisons of the four schemes are summarised in table 4.7 in which the DR penalty relative to rectangular PDF dither is shown. A negative number indicates that the scheme has a greater DR penalty the rectangular PDF dither. The following observations can be made:

- Triangular PDF dither has poorer performance than the other three schemes, especially for wideband linearisation. This is because high levels of triangular dither seriously degrade the stability of the modulator. This is demonstrated by the low values of P_m in table 4.4.

- For third order modulation, DBF is unsuccessful at linearising the modulator and 1-bit dither can only remove baseband tones.

Order	Baseband			Wideband		
	Tri PDF	1-bit	DBF	Tri PDF	1-bit	DBF
3	-0.9	1.2	-	-3.0	-	-
4	-0.8	-0.5	1.1	-4.2	-2.3	3.9
5	-2.8	-1.7	-0.6	-3.9	3.9	6.4
6	-0.8	-0.7	-0.3	-7.7	-0.1	4.0

Table 4.7: Noise Performance Relative to Rectangular PDF. All units in *dB*

- For fourth and higher order modulation, DBF offers slightly lower noise penalties than 1-bit dither for baseband linearity. Its performance relative to rectangular PDF dither is order dependent. In both cases the differences are within 1.1 *dB*.

- For fourth and higher order modulation, DBF offers significantly lower noise penalties for wideband linearity.

To complete the comparison, the noise modulation of each scheme has been evaluated over the input range -120 *dB* to -20 *dB* with a DC increment of 0.5 *dB*. The top 20 *dB* of the dynamic range has not been evaluated, since overloading can cause the noise power to increase significantly as maximum input level is reached. The peak variation in noise power is shown in tables 4.8 and 4.9 for baseband and wideband tone attenuation, respectively. Where the label 'instab' is shown, the modulator exhibits unstable regions for some of the DC inputs within the normally stable amplitude range. This behaviour is termed *unreliable* operation [Ris94b]. This occurs only for third order modulation and indicates that the dither causes the modulator to enter unstable regions in the state space which are not normally entered by undithered modulators. The issue of reliability is beyond the scope of this thesis, however the reader is referred to [Ris94b] and [Ker96] for more information.

For baseband tone attenuation, the peak noise modulation is below 0.6 *dB* for all orders and dithering schemes (with the exception of third order DBF which has not been measured due to its inability to remove baseband tones). In [Dun94] it is stated that about 1 *dB* of noise modulation can be detected under critical listening conditions if the noise floor is above the threshold of audibility. Therefore, according to this criterion, the level of noise modulation in table 4.7 will not be audible. It is likely that noise modulation occurring at high input levels (above -20 dB) would

Order	Rectangular	Tri PDF	1-bit	DBF
3	0.39	instab.	0.45	-
4	0.36	0.39	0.38	0.39
5	0.41	0.36	0.56	0.41
6	0.47	0.39	0.41	0.34

Table 4.8: Noise Modulation in dB for Baseband Tone Attenuation

Order	Rectangular	Tri PDF	1-bit	DBF
3	instab.	instab.	0.44	-
4	0.37	0.41	0.44	1.76
5	0.38	0.41	0.37	1.17
6	0.31	0.39	0.51	0.94

Table 4.9: Noise Modulation in dB for Wideband Tone Attenuation

also be inaudible due to signal masking [Nor95].

For wideband tone attenuation, the noise modulation is kept below 0.51 dB, with the exception of DBF. The higher noise modulation with DBF is due to an unusual characteristic, in which the noise power *reduces* at high input levels. An example of this is shown in figure 4.29 for the fifth order modulator with DBF. The reduction in noise power is due to the bit-flipping algorithm beginning to turn off at high input levels. It has been shown in chapter 3 that at high input levels, a reduction in quantizer gain occurs and this corresponds to an increase in variance of $u(k)$. As a consequence the proportion of samples which satisfy the bit flipping criterion $|u(k)| < B$ reduces, and there is an associated reduction in noise penalty.

4.6 Summary

This chapter has focussed on the use of bit-flipping to linearise $\Sigma\Delta$ modulators. A model of a dithered modulator has been proposed, in which the dithered quantizer is replaced by a quantizer with bit-flipping. In the model, the dither occasionally changes the output state of the quantizer, and this has the effect of breaking up limit cycles in the one-bit output and causing the idle tones to be attenuated. A fundamental proposition is made that it is possible to implement dither directly as a bit-flipping operation. It has been shown that with knowledge of the limit

cycle composition of the output, it is possible to disrupt attracting limit cycles by appropriate bit-flipping. Two problems have been identified with this scheme. The most fundamental is that it is unrealistic to detect and disrupt all possible limit cycles which cause tones. Secondly, bit-flipping causes the quantizer error magnitude to increase and this can lead to instability.

An alternative linearisation scheme has been investigated which emulates dither. A simple condition has been obtained under which dither causes a change in output state. This condition has been used to obtain a bit-flipping condition which has no random element i.e. the linearisation is deterministic. The condition also ensures that the maximum quantizer error is bounded for samples in which bit-flipping occurs and this enhances the stability of the system. A control parameter B, termed the quantizer input bound, controls bit-flipping activity. Simulations have shown that this deterministic bit-flipping (DBF) is capable of linearising the modulator by attenuating idle tones and reducing noise modulation. The attenuation of the dominant low and high frequency tones varies periodically with B. For fourth and higher order modulators it is possible to choose B to simultaneously attenuate both low and high frequency tones.

The quasi-linear model has been extended to model the noise performance of rectangular PDF dithered modulators and DBF modulators. The quasi-linear parameters are obtained by solving a pair of nonlinear equations in two unknowns. The solutions have shown that an increase in the dither level (for the dithered modulator) or quantizer input bound (for the DBF modulator), causes the quantizer noise gain K_n to reduce and the error variance σ_e^2 to increase. As a consequence the baseband noise power increases and the MSA reduces. A duality between dither and DBF has also been identified which shows that in terms of the quasi-linear model, dither and DBF have an almost identical effect on the behaviour of the modulator for an appropriate scaling between the dither level and quantizer input bound.

A comparison of the performance of different linearisation schemes has been performed including rectangular dither, triangular high-pass dither, 1-bit dither and DBF. The aim has been to establish the dynamic range penalty incurred in achieving both baseband and wideband linearity. The linearity has been measured in terms of the level of baseband and high frequency tones relative to the noise floor, followed by noise modulation plots to verify the results.

The results show that triangular high-pass dither has the highest dynamic range

penalties. The DBF technique is capable of attenuating tones for third and higher order modulation, however complete linearisation is not possible in the third order case. For fourth and higher order modulators, DBF has a slightly lower dynamic range penalty than 1-bit dither for baseband linearisation. Its performance relative to rectangular PDF dither is order dependent, with slightly inferior performance obtained for 5th and 6th order modulators. For wideband linearisation, DBF offers significantly lower dynamic range penalties than all the other schemes. Additionally, its implementation is efficient and is well suited to A-D conversion implementation.

Chapter 5

Adaptive Bit-Flipping Architectures

5.1 Introduction

It has been shown in chapter 4 how bit-flipping may be used to enhance the linearity of the $\Sigma\Delta$ modulator. In this chapter the following question is addressed: Is it possible to use bit-flipping to enhance the dynamic range of the modulator? Using standard (fixed) architectures, the dynamic range is governed by the power gain and baseband attenuation of the noise transfer function. It has been shown in chapter 2 that there is a tradeoff between these two properties. Referring to the Noise Shaping Theorem (theorem 2.2), reducing the power gain of the NTF to accommodate a greater signal headroom, requires that the transition width of the filter is reduced. To achieve this the out-of-band poles must be closer to the baseband edge and as a consequence the NTF baseband attenuation suffers and the baseband noise power increases. There are two ways in which the tradeoff can be improved. Either the order of the filter must be increased, resulting in an increase in modulator complexity; or the modulator can be made adaptive. The concept of adapting either the quantizer or loop filter has been introduced in section 2.5. In this chapter bit-flipping algorithms are investigated which have the effect of adapting the noise spectrum of the modulator in a manner which is dependent on the input level. The investigations begin with the following proposition.

Proposition 5.1 *A high order modulator (either adaptive or fixed) can be emulated using a low order modulator with bit-flipping.*

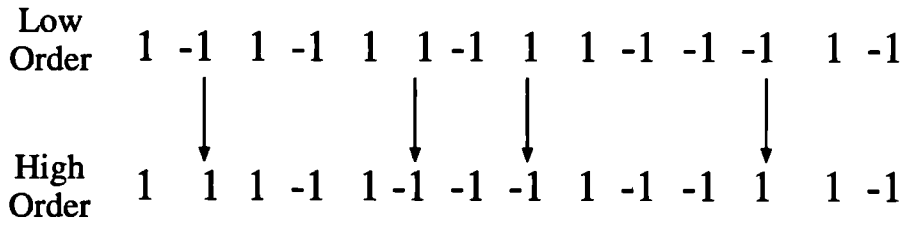


Figure 5.1: Conceptual difference between low and high order modulator outputs

This is based upon the observation that outputs from low and high order modulators differ only by a sequence of selective output inversions [San95] (figure 5.1).

A further motivation for testing the hypothesis is to show from a theoretical viewpoint that, unlike random dither, bit-flipping does not *inherently* cause the stability and baseband noise performance to suffer. The models used to explain the increase in baseband noise in chapter 4 have made the fundamental assumption that successive error samples generated by the bit-flipping or dithered quantizer are uncorrelated with each other i.e. the error is spectrally white.

It is the initial aim of this chapter to demonstrate how the combined quantizer and bit-flipping error may be auto-correlated, enabling its spectral response to be shaped as desired and therefore to test Proposition 5.1. This dissertation is concerned only with baseband applications of $\Sigma\Delta$ modulation and so the study is restricted to the generation of a high pass quantization error, allowing the quantization noise to be attenuated in the baseband over and above the attenuation provided by the noise shaping.

5.2 Weighted Bit Flipping Algorithm

In this section a bit-flipping algorithm is developed which causes the baseband quantization error power to be reduced. The algorithm hinges on the estimation of the baseband quantizer error variance through a weighting filter $W(z)$ which emphasises baseband frequencies (figure 5.2). For an assumed zero mean quantization error sequence denoted in the z-domain by the additive signal $E(z)$, the error variance measured through the weighting filter is given by:

$$\sigma_w^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} |E(\theta)|^2 |W(\theta)|^2 d\theta \tag{5.1}$$

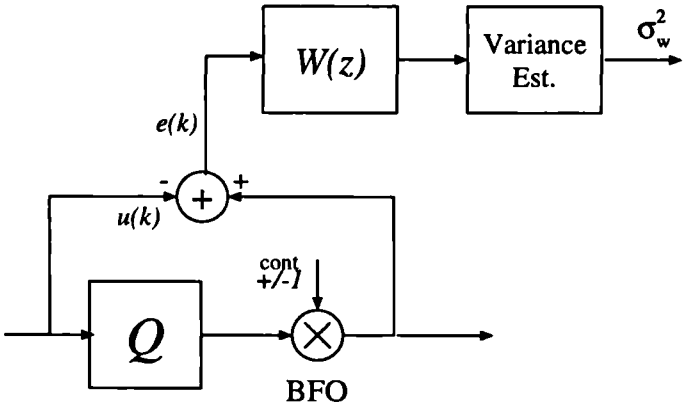


Figure 5.2: Estimation of Baseband Quantization Error Power through Weighting Filter.

By minimising σ_w^2 the AC power in the band which is emphasised by the weighting filter will be reduced. Denoting $E_w(z) = E(z)W(z)$, σ_w^2 can be expressed in the time domain, in terms of the inverse z-transform of $E_w(z)$:

$$\sigma_w^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} e_w^2(k) \tag{5.2}$$

To attenuate the baseband noise power, the bit-flipping operator (BFO) must be controlled in such a way as to minimise σ_w^2, and this is shown conceptually in figure 5.3. The decision whether to activate the BFO for a particular sample is governed by two conditions which are derived below: a *high-pass condition*, which establishes which are the best samples to flip to minimise σ_w^2, and a *stability condition*, to ensure that the samples which are flipped have a minimum impact on the stability of the modulator. The latter condition is required because bit-flipping can degrade stability (refer to chapter 4). The technique is termed weighted bit-flipping (WBF).

The simplest possible weighting filter is the first order discrete-time integrator $W(z) = z/(z-1)$ which has a pole at DC and therefore has a high gain at low frequencies. This filter is used in the following because it results in the simplest possible WBF algorithm and so allows its fundamental characteristics to be identified. Later, more complex filters will be investigated.

5.2.1 High-pass Condition

The high-pass condition is designed to test which output samples need to be inverted to minimise the weighted error variance. The derivation begins with a comparison

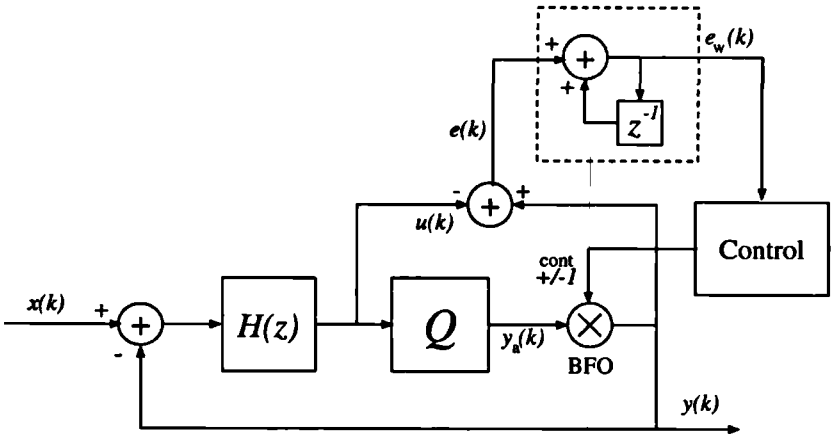


Figure 5.3: Conceptual Block diagram of $\Sigma\Delta$ modulator with Weighted Bit Flipping

of the variance of the weighted error $\sigma_w^2(k)$ with and without the inversion of the k^{th} (current) sample. An estimate of σ_w^2 after k samples is given by:

$$\sigma_w^2(k) = \frac{1}{k}\sum_{j=1}^{k} e_w^2(j) = \frac{1}{k}\left\{\sum_{j=1}^{k-1} e_w^2(j) + e_w^2(k)\right\} \tag{5.3}$$

The signals are defined in figure 5.3.

$$\begin{aligned} e_w(k) &= e_w(k-1) + e(k) \\ &= e_w(k-1) + y(k) - u(k) \end{aligned} \tag{5.4}$$

$$y(k) = \begin{cases} y_a(k) & \text{no flipping} \\ -y_a(k) & \text{with flipping} \end{cases} \tag{5.5}$$

The weighted variance without flipping is:

$$\sigma_{wn}^2(k) = \frac{1}{k}\left\{\sum_{j=1}^{k-1} e_w^2(j) + (e_w(k-1) - u(k) + y_a(k))^2\right\} \tag{5.6}$$

and with flipping:

$$\sigma_{wf}^2(k) = \frac{1}{k}\left\{\sum_{j=1}^{k-1} e_w^2(j) + (e_w(k-1) - u(k) - y_a(k))^2\right\} \tag{5.7}$$

To minimise the variance the decision is to flip if $\sigma_{wf}^2(k) < \sigma_{wn}^2(k)$, that is:

$$(e_w(k-1) - u(k) - y_a(k))^2 < (e_w(k-1) - u(k) + y_a(k))^2 \tag{5.8}$$

leading to the high pass condition:

$$|e_w(k-1) - u(k) - y_a(k)| < |e_w(k-1) - u(k) + y_a(k)| \tag{5.9}$$

The derivation shows that the minimisation of σ_w^2 is achieved when the quantizer output sample $y_a(k)$ is inverted for all samples which satisfy condition 5.9.

5.2.2 Stability Condition

In section 4.3.3 it was shown that bit-flipping can degrade the stability of the modulator due to a growth in the quantizer error. The extent to which this occurs depends on the magnitude of the quantizer error for the samples which are flipped. A useful characteristic of the deterministic bit flipping technique of chapter 4 is that when bit-flipping occurs, the instantaneous quantization error is bounded. This reduces the likelihood of quantizer overload which may trigger instability. Recalling from section 4.3.3, for a quantizer input magnitude $u(k)$ and quantizer levels ± 1, the instantaneous quantizer error with bit-flipping is equal to $1 + |u(k)|$. Therefore by only flipping when $|u(k)| < B$, the quantizer error with bit-flipping will be bounded by $1 + B$. It is possible to use the same bounding method in WBF, resulting on the stability condition:

$$|u(k)| < B \tag{5.10}$$

By restricting the quantizer inversion to samples which satisfy this condition, the stability of the modulator can be controlled by choice of the constant B.

The operation of the WBF algorithm with both the high-pass and stability condition active is:

Invert the quantizer state if both conditions 5.9 and 5.10 are satisfied.

It is clear that the stability condition will constrain the minimisation of σ_w^2 by preventing the inversion of some samples which satisfy the high pass condition. Therefore B also acts as a control parameter which governs the the extent to which the algorithm is turned on. With $B = 0$ the algorithm is fully off and the operation is identical to a fixed modulator with loop filter $H(z)$.

5.2.3 Examples of WBF

Results are now presented of simulations of the WBF algorithm with a first order weighting filter for the modulator {64, 4, 3.0} with a 1 kHz sinusoidal input of amplitude A_s. The notation $\{L, N, P_n\}$ is used to define the oversampling ratio, order and power gain in exactly the same way as in previous chapters (refer to appendix A.2). The order N and power gain P_n refer to the noise transfer function $1/(1 + H(z))$. Figure 5.4 shows the baseband results of two simulations with parameters $A_s = 0.2$, $B = 0$ and $A_s = 0.2$, $B = 0.3$. This confirms that the WBF algorithm is capable of reducing the baseband noise power of the modulator. Due to the DC pole of the weighting filter, the minimisation of the noise power is most effective at DC.

The turn-on characteristic of the WBF algorithm for the above modulator parameters as B increases from zero is shown in figure 5.5(a). For low levels of B, the baseband noise power P_b increases with B. Above a certain level (here $B = 0.05$), P_b falls with B to a level approximately 22 dB lower than for $B = 0$. The initial increase in P_b as the algorithm turns on is investigated in detail in appendix D.1.

In figure 5.5(b), the effect of varying the input level A_s for a modulator with parameters {64, 4, 1.0} and fixed B is shown. Results are shown for the three values: $B = 0$, $B = 0.2$, $B = 0.4$. The graphs presented are the results of *isolated* simulations i.e. each point on the graphs represents a new simulation. This means that they represent steady-state, not dynamic performance. These results illustrate that the baseband noise power is highly signal dependent. As the input amplitude increases the noise power rises sharply to a level greater than with the fixed modulator. The input amplitude at which the noise power begins to rise increases with B.

5.3 Modelling Weighted Bit-Flipping

Section 5.2.3 showed that the quantization noise spectrum can be shaped using a bit-flipping operation on the quantizer output. The attenuation of baseband noise confirms that higher order modulation can be emulated using bit-flipping, and so Proposition 5.1 is confirmed.

It is the purpose of this section to analyse the WBF system in order to reveal its fundamental operating mechanisms. An equivalent system is derived, in which

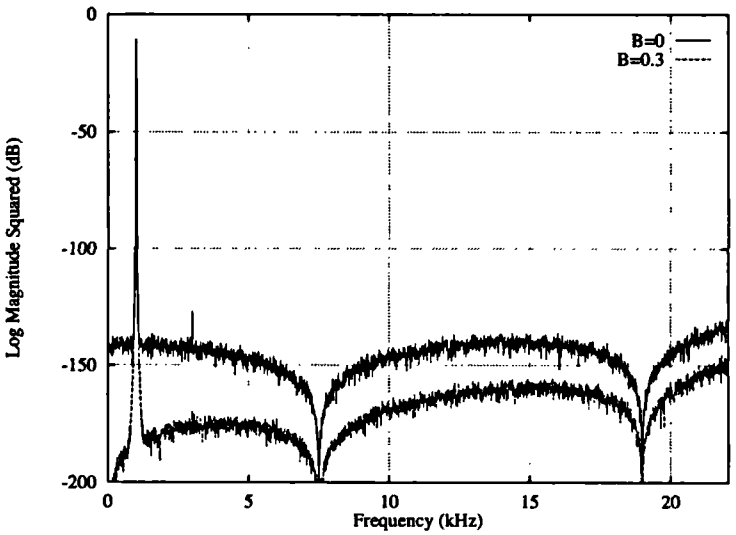


Figure 5.4: Baseband FFT of $\Sigma\Delta$ modulator without bit-flipping and with bit-flipping ($B = 0.3$) for parameters {64, 4, 3.0}

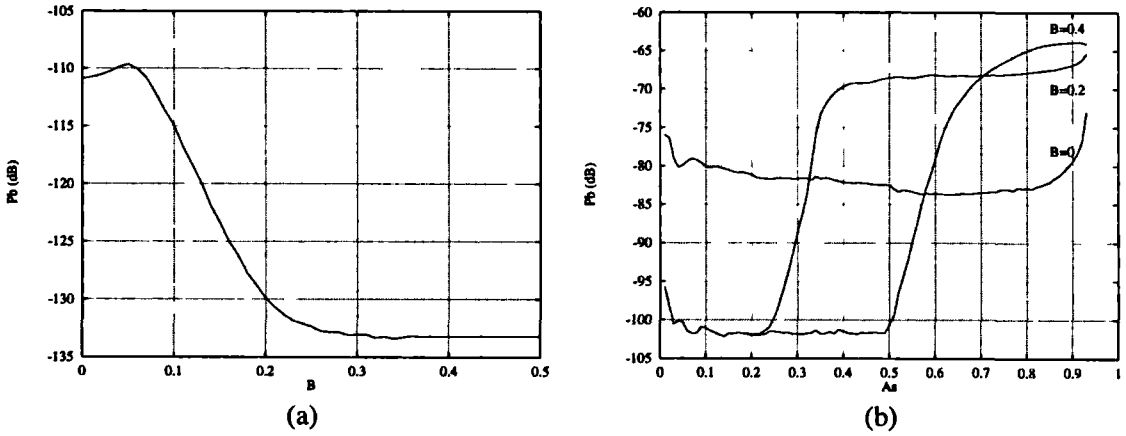


Figure 5.5: (a) Variation of Baseband Noise power with B for modulator {64, 4, 3.0} with $A_s = 0.2$, (b) Variation of Baseband noise power with A_s for modulator {64, 4, 1.0}.

the bit-flipping operation is modelled as an equivalent operation on the NTF. The model will be used later to extend the technique to use higher order weighting filters and guide in the design of some example systems.

To simplify the derivation of the equivalent system, the stability condition 5.10 is first removed. This allows the high-pass condition 5.9 to be successively broken down into simpler operations, which can then be restructured to form the equivalent system.

Starting with the high pass condition, the substitution is made:

$$v(k) = u(k) - e_w(k-1) \tag{5.11}$$

Therefore condition 5.9 can be re-written as:

$$|-v(k) - y_a(k)| < |-v(k) + y_a(k)| \tag{5.12}$$

Because the quantizer has only two states, this can be rewritten as two separate cases. Each case is directly equivalent to a simpler condition:

- Case 1: $y_a(k) = +1$

$$|-v(k) - 1| < |-v(k) + 1| \;\Rightarrow\; v(k) < 0 \tag{5.13}$$

- Case 2: $y_a(k) = -1$

$$|-v(k) + 1| < |-v(k) - 1| \;\Rightarrow\; v(k) > 0 \tag{5.14}$$

Both cases can be combined, reducing condition 5.9 to the inequality test.[1]

$$sgn\{v(k)\} \neq y_a \tag{5.15}$$

Noting that $y_a = sgn\{u(k)\}$, the combined operation of the quantizer and BFO with the stability condition inactive can be written as:

$$y(k) = \begin{cases} -sgn\{u(k)\} & \text{if } sgn\{v(k)\} \neq sgn\{u(k)\} \\ +sgn\{u(k)\} & \text{otherwise} \end{cases} \tag{5.16}$$

By considering the possible signs of $u(k)$ and $v(k)$ (table 5.1) equation 5.16 reduces to:

$$y(k) = sgn\{v(k)\} \tag{5.17}$$

$sgn\{u(k)\}$	$sgn\{v(k)\}$	$y(k)$
+	+	+
+	-	-
-	+	+
-	-	-

Table 5.1: Simplification of expression 5.16.

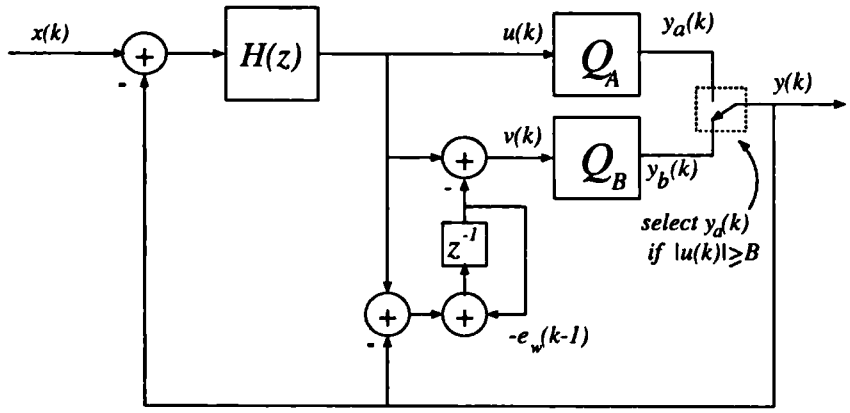


Figure 5.6: Dual Quantizer Model of $\Sigma\Delta$ Modulator with Weighted Bit Flipping

Finally the stability condition is introduced again:

$$y(k) = \begin{cases} sgn\{v(k)\} & \text{if } |u(k)| < B \\ sgn\{u(k)\} & \text{if } |u(k)| \geq B \end{cases} \tag{5.18}$$

The block diagram of the resulting equivalent system is shown in figure 5.6.

This system now has two quantizers, Q_A and Q_B, and a selector which chooses the output and feedback connection on a sample-by-sample basis, depending on the stability condition, as given by equation 5.18. This model is termed the *dual quantizer model.*

[1] A special case arises for $v(k) = 0$ in which the high-pass condition is not satisfied for any $y_a(k)$. For this case, the resulting model is imprecise, however, simulations have shown that the practical occurrence of this condition is extremely rare and the accuracy of the model is unaffected.

5.3.1 Fixed System Analysis

A starting point in the analysis of the dual quantizer model is to consider the two extreme cases where either Q_A or Q_B is permanently selected (i.e. the modulator is 'fixed'). By replacing the two quantizers with additive error sequences defined exclusively as $E_a(z)$ and $E_b(z)$ respectively, it is possible to analyse the system in the z-domain.

1. Q_A permanently selected. The modulator output is given by:

$$Y_a(z) = X(z)STF(z) + E_a(z)NTF_a(z) \tag{5.19}$$

where

$$STF(z) = \frac{H(z)}{1+H(z)} \tag{5.20}$$

$$NTF_a(z) = \frac{1}{1+H(z)} \tag{5.21}$$

The modulator operation is identical to a fixed modulator with loop filter $H(z)$.

2. Q_B permanently selected. The system is equivalent to the WBF modulator with the stability condition inactive i.e. the high-pass condition alone is operating. The modulator output is found by:

$$\begin{aligned} U(z) &= (X(z) - Y_b(z))H(z) \\ Y_b(z) &= (U(z) - Y_b(z))\frac{z^{-1}}{1-z^{-1}} + U(z) + E(z) \\ &= X(z)STF(z) + E_b(z)NTF_b(z) \end{aligned} \tag{5.22}$$

where

$$NTF_b(z) = \frac{1-z^{-1}}{1+H(z)} \tag{5.23}$$

The analysis reveals that with Q_B selected, the STF is unaltered but the NTF is modified by the $1-z^{-1}$ term. Therefore in the case where the stability condition is inactive, the bit-flipping operation can be mapped onto an equivalent operation on the NTF, and the system acts as a fixed modulator with NTF

given by equation 5.23. Due to the addition of a DC zero, the noise shaping properties of the modulator are enhanced at low frequencies. Structurally, the modulator uses local feedback around the quantizer to introduce the $1 - z^{-1}$ term. This method is similar to that used by Candy [Can85] to convert a first order modulator into a second order modulator.

5.3.2 Adaptive Control

In practical operation, the selector switch is not fixed, but selects one of the quantizers on a sample-by-sample basis. In section 5.3.3, the operation of the loop will be analysed in some detail, however for the moment, the factors which govern the operation of the selector and the resulting NTF characteristics will be discussed.

In the following, we define that by default the quantizer selector is positioned such that $y(n) = y_b(n)$ i.e. quantizer Q_B is selected and the modulator has enhanced baseband attenuation according to equation 5.23. The quantizer selector is controlled by the stability condition $|u(k)| < B$, therefore the probability density function (PDF) of $u(k)$ determines the probability of quantizer Q_A being selected:

$$p\{y(k) = y_a(k)\} = p\{|u(k)| \geq B\} \tag{5.24}$$

This probability depends on both the input signal, the modulator NTF and the degree of loop stability.

Insight into the statistical properties of $u(k)$ can be gained from the quasi-linear model. It has been shown in section 3.4 that as the input signal to a fixed modulator increases and overload is approached, the quantizer gain reduces. This indicates that the variance of $u(k)$ increases and therefore $p\{y(k) = y_a(k)\}$ increases as overload is approached.

It is possible to choose B so that at low input levels, $p\{u(k) \geq B\}$ is low, causing the modulator behaviour to be dominated by the characteristics of $NTF_b(z)$ and enhanced noise shaping to be exhibited. As the input level increases, $p\{|u(k)| \geq B\}$ increases and eventually $NTF_a(z)$ dominates. In this way the NTF adapts according to the input signal level. Since the adaption is on a sample-by-sample basis it is classified as instantaneous adaption. The adaption operates so that at high input levels, as the modulator approaches overload, the lower power gain $NTF_a(k)$ is chosen by the selector, and the MSA is increased at the expense of a deterioration in noise performance at high input levels. Therefore it is possible to view the modulator

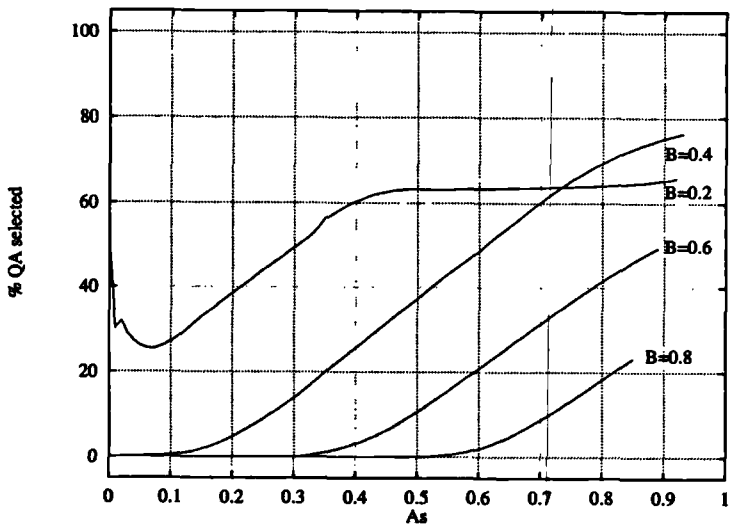


Figure 5.7: Variation of selection rate of $y_a(k)$ with A_s for the modulator {64, 4, 1.0}

from two perspectives. It can either be viewed as a modulator which uses bit-flipping to enhance the attenuation of the noise shaping at low input levels; or it can be viewed as a modulator which adapts the NTF to prevent overload at high levels and so enhances the dynamic range.

The threshold on the input level at which $p\{y(k) = y_a(k)\}$ increases is governed by the quantizer input bound, B, and this is demonstrated in figure 5.7, in which the percentage of samples where $y_a(k)$ is selected is plotted against input level for different values of B for the modulator {64, 4, 1.0}. The threshold value increases with B, though the correspondence between this value and the threshold on the noise power curves of figure 5.5 is weak. For example, with $B = 0.2$, the percentage of samples of Q_A selected is greater than zero at all input levels, yet the noise power curve has a threshold at $A_s = 0.2$. The reason for this discrepancy will be explained in section 5.3.3.

Relationship to Fixed and Adaptive Modulator Structures

It is instructive to compare a fixed modulator optimised for high input signal levels with the WBF modulator. To achieve stability at high input levels with a fixed modulator, a low power gain is required in the NTF. This limits the NTF attenuation and the noise performance obtainable since the transition band needs to be steeper. Therefore the dynamic range of the modulator suffers. The WBF modulator, on

the other hand, adapts the NTF to allow high input levels with potentially no noise penalty at low input levels. Therefore a wider dynamic range is possible than the fixed modulator. At high input levels the NTF adaption causes the baseband noise power of the modulator to increase. In an audio context, however, noise at high input levels is less critical than at low input levels due to psychoacoustic signal masking [Fie89].

It is also worthwhile commenting on the similarities between the WBF modulator and the technique of instantaneous NTF coefficient adaption in [Yu92]. Referring to section 2.5.1, the modulator coefficients are adapted according to the input signal level by means of a calibrated look-up table. The adaption enhances the dynamic range of the system. A disadvantage of this scheme is that the performance of the modulator depends on the accuracy of the calibration and so any undetected unstable regions can cause modulator instability at intermediate signal levels [Dun96].

In addition to requiring no calibration, the WBF system offers the advantage of 'automatic' stability detection. Because the NTF is directly adapted according to the degree of overload in the modulator, unstable regions at intermediate input levels can be avoided by choosing more output samples from the NTF with lower power gain. Conceptually, this is similar to the clipping scheme discussed in section 2.5.1, where filter sections are removed by clipping or reset if overload occurs.

5.3.3 Operating Regions

The concept of quantizer selection has proven useful for analysing the operation of the modulator with the selector fixed, and providing an intuitive understanding of the NTF adaption. The model does not, however, lead to a precise correspondence between the noise curves and the selection rates. Furthermore, an explanation is required for the behaviour at high input level, where referring to figure 5.5, the noise power increases to a level higher that obtained using the fixed lower order modulator obtained by choosing $B = 0$. The approach taken here in analysing the modulator is to identify operating regions which occur for different input levels by considering in greater detail the effect of the quantizer selection on the behaviour of the modulator.

Latent Region

It has been explained in section 5.3.2 that the modulator adapts by choosing Q_A when $|u(k)| \geq B$. It is clear that this selection will not affect the modulator behaviour if the two quantizer outputs are identical. This leads to the following definition:

Definition 5.1 (Latent Selection) *The quantizer selection is defined as* latent *if* $y_a(k) = y_b(k)$ *when* $|u(k)| \geq B$.

If all selections are latent for a given steady-state input, the modulator is effectively fixed with noise transfer function $NTF_b(z)$.

From definition 5.1 the condition for latency in all samples is:

$$max||u(k)||_{y_a(k) \neq y_b(k)} < B \tag{5.25}$$

This condition states: the maximum value of $|u(k)|$ measured at all sample instances where $y_a(k) \neq y_b(k)$ is less that B. This means that in the latent mode of operation $y_a(k) = y_b(k)$ for all samples in which quantizer Q_A is selected, therefore the modulator is fixed and its behaviour is determined solely by $NTF_b(z)$.

Condition 5.25 can also be interpreted in terms of the original WBF architecture of figure 5.3. By equation 5.15, $y_a(k) \neq y_b(k)$ when the high-pass condition is satisfied. Therefore, under latent operation, $|u(k)| < B$ for all samples in which the high-pass condition is satisfied i.e. bit-flipping occurs whenever the high-pass condition is satisfied.

A condition that is required for latency in all samples is that the modulator with noise transfer function $NTF_b(z)$ (or equivalently the WBF modulator with $B = \infty$) is stable. If the modulator is *unstable*, the amplitude of $|u(k)|$ will grow with k until the condition $|u(k)| \geq B$ is true for some of the instances in which $y_a(k) \neq y_b(k)$. This implies latency will occur only for small input signals below the threshold of instability.

Example of Latent Region Simulations have shown that regions of latency can exist for a range of input levels. An example of such a latent region is shown in figure 5.8 for the modulator {64, 4, 1.0} with values $B = 0.2$ and $B = 0.4$. In this figure the percentage of samples in which quantizer Q_A is actively selected is plotted against input level A_s for a 1 kHz sinewave input. This percentage will be

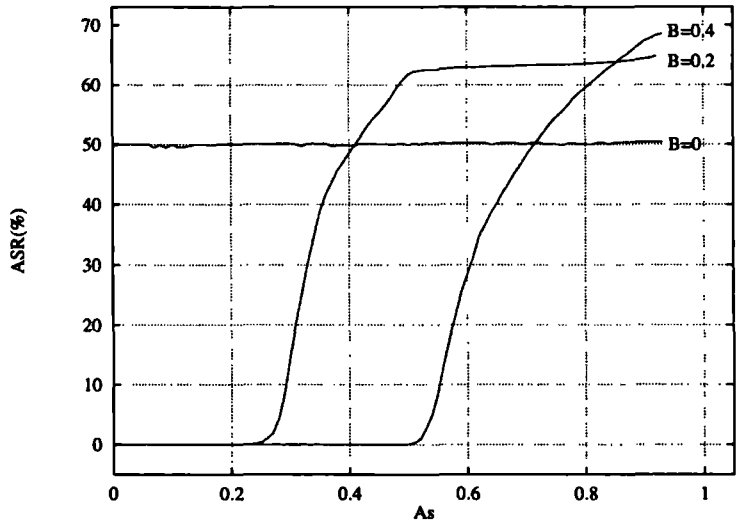


Figure 5.8: Active selection rate against A_s for the modulator {64, 4, 1.0}

referred to in the following as the active selection rate (ASR). By active selection we mean that $y_a(k) \neq y_b(k)$ whenever Q_A is selected. For each value of B in the graph, there is a threshold level, below which the active selection rate (ASR) is zero. The threshold level increases with B because the magnitude of $|u(k)|$ required to satisfy the condition $|u(k)| \geq B$ is greater. The range of inputs for which the active selection rate is zero is by definition the latent region. The modulator is fixed with noise transfer function $NTF_b(z)$ in this region. Note that there is a close correspondence between this curve and the baseband noise power curve of figure 5.5(b) which has been simulated with the same parameters. In the latent region the noise power varies very little with input amplitude since the modulator is fixed and is not overloading.

Transition and Overload Region

From figures 5.5 and 5.8, it is clear that above the latent region the ASR and the baseband noise power P_b increase rapidly with A_s. This region is termed the transition region. Above the transition region, the variation in ASR and P_b with A_s is smaller and this is termed the overload region. The three regions of operation are shown in figure 5.9.

To explain the behaviour of the modulator in the transition and overload regions it is convenient to isolate the low-order modulator in the model of figure 5.6. This modulator is shown in heavy type in figure 5.10 and is termed modulator 'A'.

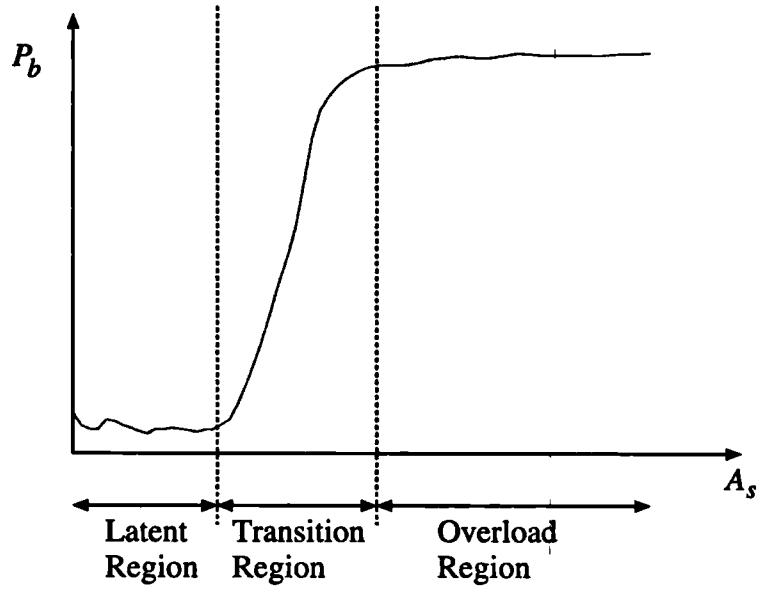


Figure 5.9: Definitions of Operating Regions for Varying Input A_s

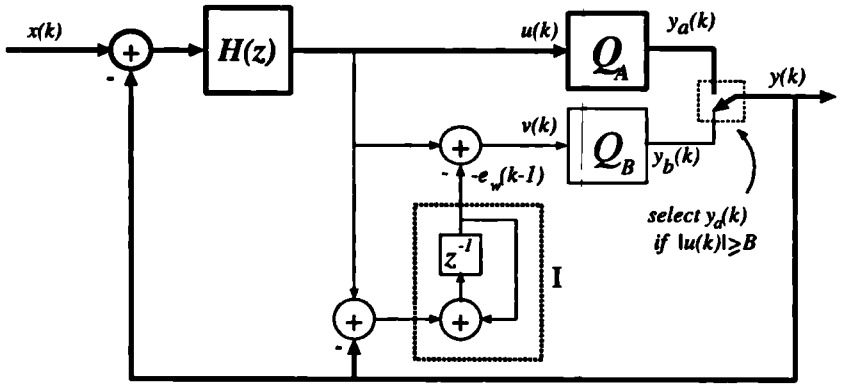


Figure 5.10: WBF Model showing low-order modulator (heavy type) and internal integrator (I)

The transition region occurs when an increase in input level causes the maximum value of $u(k)$, measured at all sample instances where $y_a(k) \neq y_b(k)$, to exceed B. In this region the NTF adapts rapidly with input level. As the selection rate of quantizer Q_A increases and the overload region is entered, quantizer Q_B becomes selected less often and, in effect, the local feedback loop opens around integrator 'I' (refer to figure 5.10). Low frequency components at the input of the integrator are amplified causing a large signal amplitude low frequency oscillation to develop in $v(k)$. As a result, $y_b(k)$ develops a pattern consisting on average a sequence of b 1's followed by b -1's (refer to appendix D.1 for further details). Under this condition the output of quantizer Q_B no longer codes the input signal, therefore the samples in which Q_B is selected represent errors to modulator A, causing the baseband noise power to increase.

The increase in signal levels at the output of the integrator may be alternatively described in terms of the WBF modulator which the model represents. At high input levels, the stability control inhibits bit-flipping for samples in which the high-pass condition is satisfied. As a result, the quantization error becomes relatively white (rather than high-pass), and so low frequency energy is amplified by the integrator, causing large signal levels to develop at its output.

5.3.4 Modulator Behaviour with Varying NTF Power Gain

In this section we consider how the operation of the WBF modulator with a first order weighting filter changes as the power gain of the $NTF_a(z)$ is varied. The power gain is denoted P_n.

In figures 5.11 to 5.14 the baseband noise power P_b and active selection rate is plotted against input level A_s for different P_n and B for a modulator with parameters $L = 64$, $N = 4$.

As the power gain increases the three operating regions become less distinct and the variations in ASR and P_b with A_s become smaller. The transition region also becomes less steep and the latent region ends at a smaller value of A_s.

There are two reasons for this behaviour. Firstly, as the power gain increases, the variance of $u(k)$ increases due to an increase in noise circulating around the loop. This is demonstrated in a fixed modulator by a decrease in quasi-linear gain (refer to section 3.4). This causes the input threshold level below which condition 5.25 is satisfied to decrease and so the latent region ends at a smaller value of A_s. It also

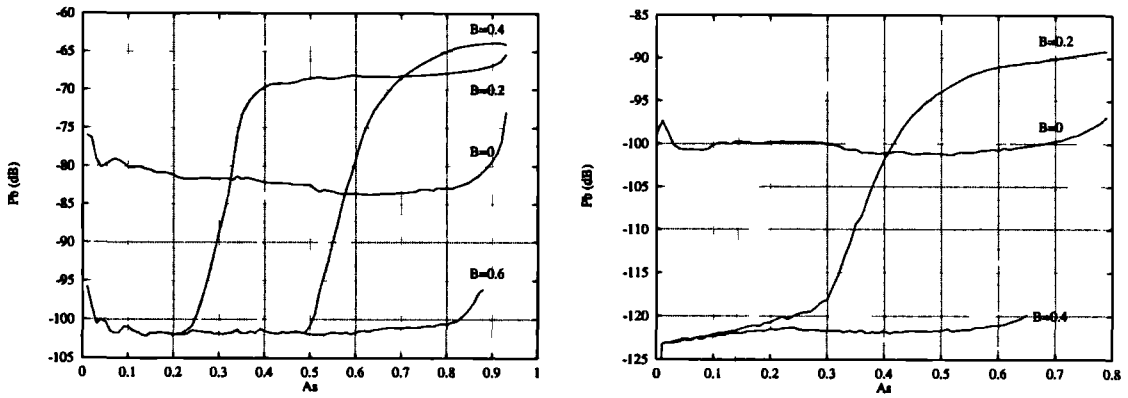


Figure 5.11: Variation of Baseband noise power with A_s for modulators (a) {64, 4, 1.0} and (b) {64, 4, 2.0}

means that $u(k)$ becomes dominated by quantization noise and the change in its variance with input level is small. Therefore the change in selection rate with input level is small and the three regions become less distinct. Secondly, the increase in P_n causes the system to become globally unstable at a smaller value of A_s due to an associated increase in the power gain of NTF_b. This can restrict the operating point from extending into the overload region.

Utilising a high P_n to avoid entry into the overload region is advantageous because it prevents excessively large wordlengths building up in the integrator adder at the input of Q_B. However as a consequence, the range of NTF adaption is relatively small. As will be seen in section 5.5 there is a compromise between the maximum allowable signal levels and dynamic range advantages gained from adaption.

B_∞-Unstable Operation

An extreme case is the use of a power gain high enough so that modulator with noise transfer function $NTF_b(z)$ is unstable at all non-zero input levels. These modulators will be termed B_∞-Unstable. Under this condition, there is no latent region and the modulator becomes adaptive over its entire operating range. It is the responsibility of the stability control to ensure that the modulator remains stable by choosing Q_A for sufficient samples. Due to the high P_n, the modulator becomes unstable well before the overload region is entered and consequently the range of adaption is small.

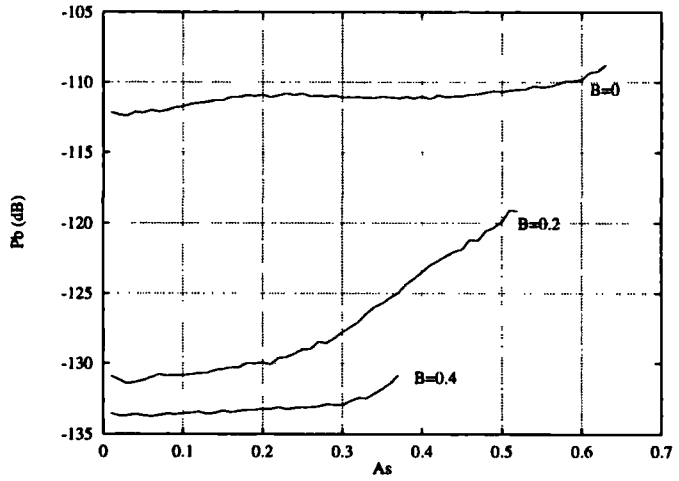


Figure 5.12: Variation of Baseband noise power with A_s for modulator {64, 4, 3.0}

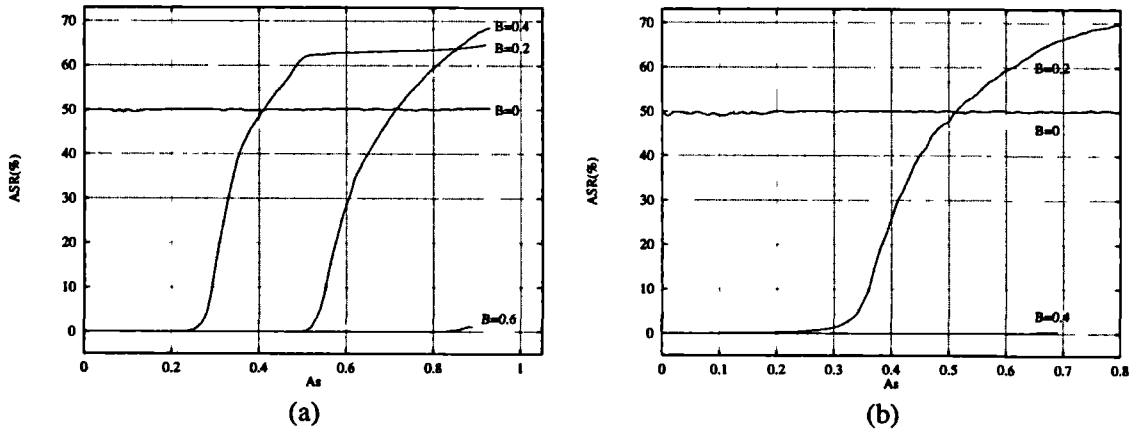


Figure 5.13: Variation of ASR with A_s for modulators (a) {64, 4, 1.0} and (b) {64, 4, 2.0}

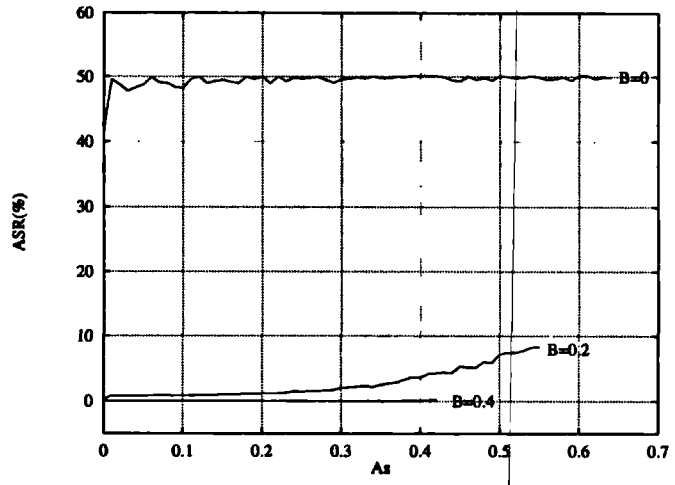


Figure 5.14: Variation of ASR with A_s for modulator {64, 4, 3.0}

5.4 Higher Order Weighting Filters

In this section it is briefly explained how the WBF technique may be extended to the use of higher order weighting filters. The motivation is to provide greater flexibility in the noise shaping properties of the modulator at low and high input levels. One example which will be investigated is the design of a system which provides better tradeoff between the low-level noise power and maximum input levels.

An obvious approach in extending the system would be to replace the first order integrator with a higher order weighting filter, which has an inverse response to the required baseband noise-shaping. It has been found, however, that difficulties arise in obtaining a simple high-pass condition which applies to a general weighting filter. An alternative method is to use the structure of the dual quantizer model of section 5.3, replacing the integrator with a filter $G(z)$. The block diagram of the system is shown in figure 5.15. Analysis reveals that the noise and signal transfer functions with quantizer Q_A and Q_B permanently selected are:

- Quantizer Q_A selected:

$$NTF_a(z) = \frac{1}{1+H(z)} \tag{5.26}$$

$$STF_a(z) = \frac{H(z)}{1+H(z)} \tag{5.27}$$

- Quantizer Q_B selected:

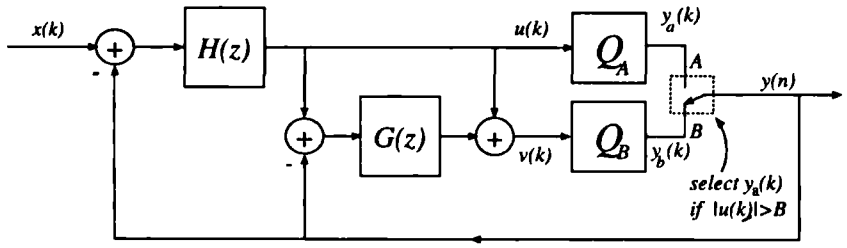


Figure 5.15: General WBF system

$$NTF_b(z) = \frac{1}{(1+H(z))(1+G(z))} \tag{5.28}$$

$$STF_b(z) = \frac{H(z)}{1+H(z)} \tag{5.29}$$

It is convenient to also define the NTF that defines the noise shaping attributable to the weighting filter:

$$NTF_w(z) = \frac{1}{1+G(z)} \tag{5.30}$$

In this study we only consider the case of FIR noise transfer functions for $NTF_w(z)$. Both $G(z)$ and $H(z)$ contain unit delays for implementation. The adder following $G(z)$, which is also present with the first order weighting filter, ensures that the signal transfer function with quantizer Q_B selected is unity, and this prevents the STF from varying as the modulator adapts.

5.4.1 Noise Transfer Function Zero Allocation

In this section we consider a NTF design strategy for general order modulators using weighted bit-flipping. An optimal design strategy would choose the NTF zeros to minimise the baseband noise power at both low and high input amplitudes.

If it is assumed that a latent region exists, the noise transfer function at low input levels in this region is given by:

$$NTF_b(z) = NTF_a(z)NTF_w(z) \tag{5.31}$$

The response at high input levels is more difficult to predict due to the overload, however it is reasonable to assume that the response is dominated by $NTF_a(z)$.

A possible design strategy is to locate the zeros of $NTF_b(z)$ to minimise the baseband noise power of the modulator as described in [Sch90]. Therefore at *low input levels* the zero locations will be optimal and the noise power minimised. The zeros of $NTF_a(z)$ are then selected from the zeros of $NTF_b(z)$ to minimise the noise power at high levels.

We denote A(z) as the zeros of $NTF_a(z)$ and $B(z)$ as the zeros of $NTF_b(z)$. We define O_a as the order of $NTF_a(z)$ and O_b as the order of $NTF_b(z)$ (note that $O_b > O_a$). For zero locations z_k which are either real or occur in complex conjugate pairs:

$$A(z) = \prod_{r=0}^{O_a-1} (z - z_r) \tag{5.32}$$

$$B(z) = \prod_{r=0}^{O_b-1} (z - z_r) \tag{5.33}$$

The set of zero locations of $B(z)$ which control the noise power at low levels are chosen from the optimal zero set of [Sch90], where the locations are defined in terms of the frequency ω_r of the zeros on the unit circle, normalised to the baseband edge:

$$z_r = e^{j\omega_r \pi/L} \tag{5.34}$$

The optimal zero frequencies are given in table 2.1 of chapter 2.

The set of zero locations of $A(z)$ which control the noise power at high levels are a subset of the zeros of $B(z)$ and are chosen to minimise the expression:

$$\int_0^{\pi/L} |A(\theta)|^2 d\theta \tag{5.35}$$

Equation 5.31 is then used to define the zeros of $NTF_w(z)$ and hence the filter $G(z)$. The technique is illustrated in the design of example systems C and E in section 5.5.

5.5 Investigations and Results

Results are now presented for some example systems. Systems A, B and C are included to show that it is possible to use weighted bit-flipping to stabilise modulators with Tewksbury NTFs of order greater than two (refer to section 2.3.2). These

modulators are normally unstable. System C demonstrates how the zero locations may be modified to enhance the dynamic range of the modulator. System D and E use an IIR filter for $NTF_a(z)$ and it is shown how the modulator may be designed to avoid overload and achieve a greater dynamic range than a fixed modulator.

Unless otherwise stated, all the results presented in the following are based upon an example modulator with the following characteristics:

- Sampling frequency $F_s = 44.1\ kHz$.
- Oversampling ratio $L = 64$.
- Input signal = 1 kHz sinewave of peak amplitude A_s.

5.5.1 System A - Second Order FIR $NTF_a(z)$, First Order FIR $NTF_w(z)$, DC zeros in NTFs.

The noise transfer functions are given by:

$$NTF_a = (1 - z^{-1})^2 \tag{5.36}$$

$$NTF_w = (1 - z^{-1})^1 \tag{5.37}$$

resulting in $NTF_b(z) = (1 - z^{-1})^3$

With $B = \infty$ (the stability condition inactive) the modulator is equivalent to a fixed 3rd order Tewksbury modulator with DC zeros, which is unstable for all non-zero inputs [Can85] i.e. the system operates in B_∞-Unstable mode and the stability control is continually active to ensure that the modulator is stable (refer to section 5.3.4).

In the design of this system, the only unknown variable is B. Since the WBF algorithm enhances the modulator performance at low input levels, the optimal value of B has been determined for a low input level of $A_s = 0.01$. In figure 5.16(a) the baseband noise power P_b is plotted against B. The noise power is minimised at a value of $B = 0.62$. In figure 5.16(b) the baseband noise power is plotted against A_s, using the optimal value $B = 0.62$. Notice that the transition region is gradual and there are no distinct latent or overload regions. This is because the modulator operates in B_∞-Unstable mode. On the same axis the noise power for a second order Tewksbury modulator, with $NTF = (1 - z^{-1})^2$ is plotted. The WBF

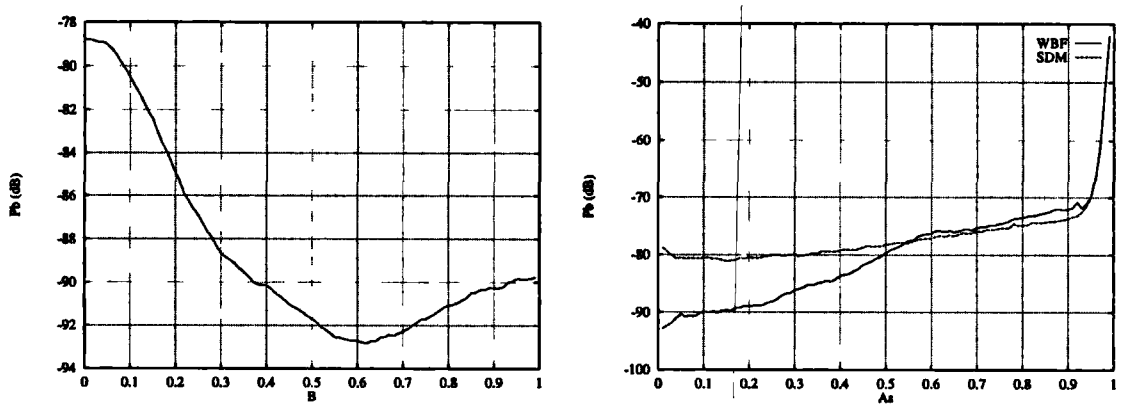


Figure 5.16: System A: Variation of P_b with (a) B and (b) A_s

system achieves a reduction in noise power compared with the fixed modulator of approximately 14 dB at $A_s = 0.01$. At higher input levels, the WBF modulator has a slightly higher noise power due to higher signal levels in $u(k)$, causing a reduction in quasi-linear gain in quantizer Q_A (refer to section 5.3.2).

Both modulators are stable for input levels up to full scale, though the noise power rises rapidly above $A_s = 0.9$ as the modulators overload. The identical maximum input level of both modulators confirms the intended adaptive operation of the WBF modulator i.e. the modulator has the stability of a second order modulator but the low-level noise power of a higher order modulator. Comparisons with a third order 'Tewksbury' modulator is not possible because this modulator is unstable with all non-zero inputs. The spectral response of the WBF system with inputs of $A_s = 0.01$ and $A_s = 0.8$ are shown in figure 5.17. The harmonic distortion with $A_s = 0.8$ occurs also for the second order modulator and is caused by the modulator approaching overload.

5.5.2 System B and C - Second Order FIR $NTF_a(z)$, Second Order FIR $NTF_w(z)$

- System B has DC zeros in the NTFs.

$$NTF_a = NTF_w = (1 - z^{-1})^2 \tag{5.38}$$

therefore $NTF_b(z) = (1 - z^{-1})^4$

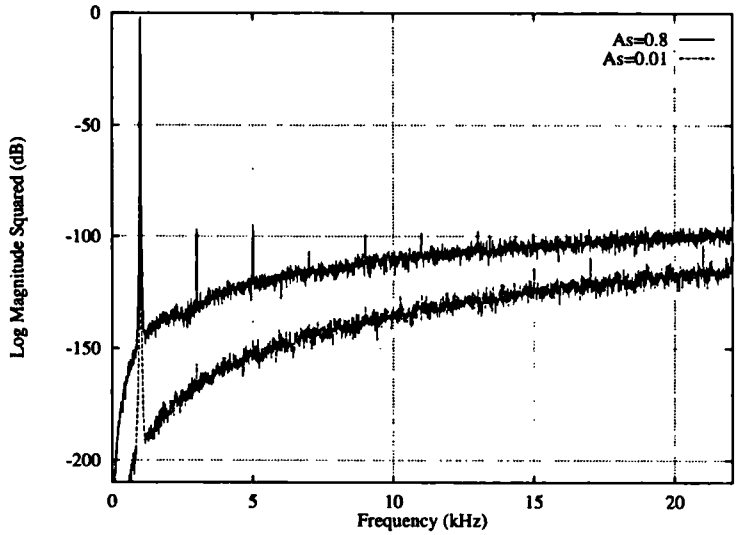


Figure 5.17: System A: Baseband spectral response with inputs $A_s = 0.01$ and $A_s = 0.8$

With $B = \infty$ the modulator is equivalent to a 4th order Tewksbury modulator with DC zeros, which is unstable for all non-zero inputs and therefore the system operates in B_∞-Unstable mode.

- System C has complex conjugate zeros in the NTF.

  This system is presented to demonstrate the improvement in performance possible by optimally locating the noise transfer function zeros across the baseband. The zeros of $NTF_b(z)$ which define the noise performance at low levels are all complex conjugate and on the unit circle given by:

$$\begin{aligned} NTF_b(z) &= NTF_a(z)NTF_w(z) \\ &= (z - e^{j\omega_a \pi/L})(z - e^{-j\omega_a \pi/L})(z - e^{j\omega_w \pi/L})(z - e^{-j\omega_w \pi/L}) \end{aligned} \tag{5.39}$$

  where ω_a and ω_w are the normalised zero frequencies of NTF_a and NTF_w respectively. From table 2.1 the two normalised optimal zero frequency pairs of a fourth order $L = 64$ system are:

$$\omega_0 = \pm 0.33998 \tag{5.40}$$

$$\omega_1 = \pm 0.861136 \tag{5.41}$$

ω_a	Noise Power
0.33998	-75.2 dB
0.86114	-80.4 dB

Table 5.2: Noise power obtained for $NTF_a(z)$ with selected zero locations.

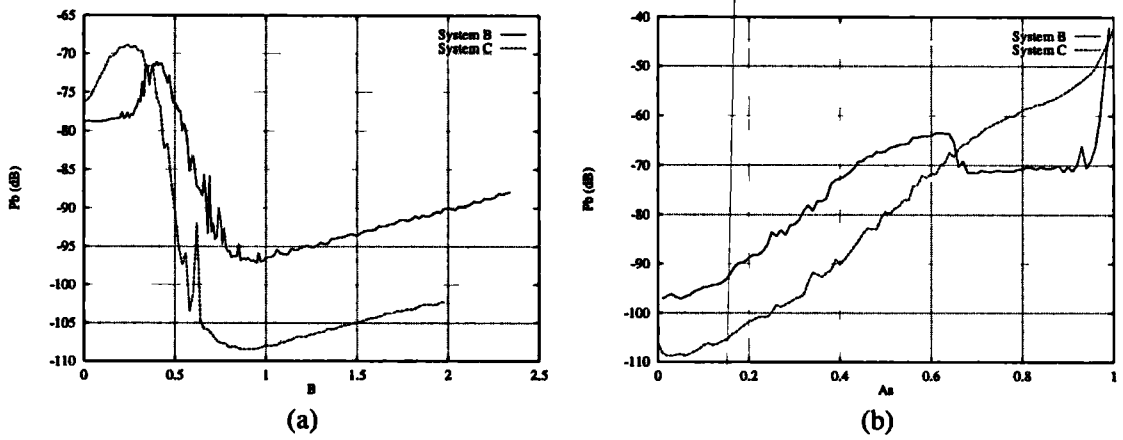


Figure 5.18: System B and C: Variation of P_b with (a) B and (b) A_s

At high input levels, where the assumption is made that $NTF_a(z)$ dominates, the complex conjugate pair which produces the lowest baseband noise power must be chosen as ω_a. To determine which pair to use, simulations of a fixed modulator have been performed with each zero pair. The baseband noise power of a second order modulator obtained with a sinusoidal input of $A_s = 0.5$ for each zero location are given in table 5.2. This method obtains the same zero locations as found by minimising the integral of equation 5.35.

ω_a is chosen as the zero pair which yields the smallest noise power, hence $\omega_a = 0.86114$ leading to the choice $\omega_w = 0.33998$.

The simulations used for system A have been repeated with system B and C. The baseband noise power, P_b, of both systems are plotted against B in figure 5.18(a). This figure clearly shows the existence of an overload and a sharp transition region (refer to appendix D.1). This is because, compared to the first order $G(z)$ used in system A, the second order $G(z)$ has greater gain at signal frequencies, making overload in $v(k)$ more likely. Neither system B or C have a latent region due to the B_∞-Unstable mode of operation. The minimum in the $P_b(B)$ curve occurs at $B = 0.95$ for both modulators.

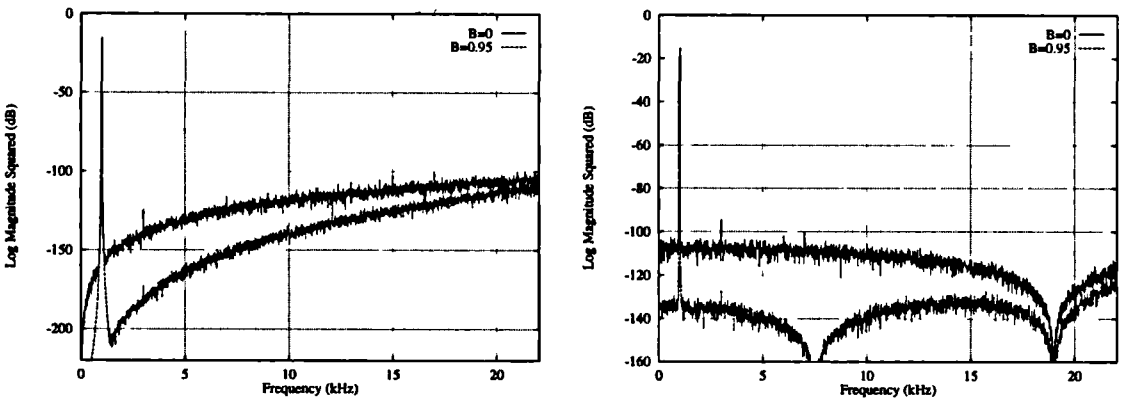


Figure 5.19: Baseband spectral response with inputs $A_s = 0.1778$ and $B = 0$ (upper), $B = 0.95$ (lower) for (a) System B and (b) System C

In figure 5.19 the baseband spectral responses of the two systems are shown with $A_s = 0.1778$ for $B = 0$ and $B = 0.95$. In system B, all the zeros are at DC, whereas in system C they are spread across the baseband. With $B = 0.95$ the modulator zeros are optimally located and the baseband noise power is approximately 11.2 dB lower than with DC zeros (figure 5.18). With $B = 0$ the system NTF ($= NTF_a(z)$) is fixed, with a complex conjugate zero pair at $\omega = 0.86114$ and this results in a larger noise power that with the zero pair at DC. Optimal performance for low-level inputs will therefore be achieved at the expense of poorer performance at high-level inputs.

This is demonstrated in figure 5.18(b), in which the baseband noise power P_b is plotted against input amplitude A_s for both systems with the optimal value $B = 0.95$. For input levels below $A_s = 0.66$, the baseband noise power of system C is lower than system B, however the reverse is true for inputs levels above $A_s = 0.66$. System C has a smoother noise power characteristic, possibly because the use of non-DC zeros in $NTF_w(z)$ restricts the maximum gain of $H(z)$ and the amount of overloading which occurs.

The performance of systems A, B and C in relation to a fixed second order modulator is summarised in table 5.3. The dynamic range is an approximation, obtained as the ratio of the maximum signal level and the noise power at $A_s = 0.01$. The assumption is made that the noise power below $A_s = 0.01$ is constant. At low input levels ($A_s = 0.01$) the performance increases in the order 2nd < A < B < C. At intermediate levels ($A_s = 0.8$) the reverse is true, due to overloading in the local

	P_b (dB)				
A_s	2nd	A	B	C	4th
0.01	-78.8	-92.8	-97.2	-108.2	-119.8
0.80	-74.9	-73.6	-70.8	-58.9	-
0.99	-42.1	-42.1	-42.1	-44.5	-
Max input	0.99	0.99	0.99	0.99	0.33
Dyn. Range (dB)	78.8	92.8	97.2	108.2	110.2

Table 5.3: Performance of System A, B, C and fixed second order modulator (- represents instability)

feedback loop. At maximum input level, the performance of all methods are similar, because the behaviour is dominated by global modulator overload. Also plotted in the table are the results for a fixed fourth order modulator with a NTF power gain $P_n = 4.2$, chosen to maximise the dynamic range. This system has greater complexity than the WBF modulator, due to the specification of poles as well as zeros, however, the dynamic range is only 2 dB greater. For a real implementation, the operation in the overload region must also be considered and this is discussed in the next section.

5.5.3 System D - Fourth Order IIR $NTF_a(z)$, First Order FIR $NTF_w(z)$.

This system is the same as the WBF system used for the examples of sections 5.2.3 to 5.3.4, with the exception that optimal zero frequencies are used.

$NTF_a(z)$ is a general IIR filter with poles in a Butterworth configuration (refer to section 2.3.1)
$NTF_w(z) = 1 - z^{-1}$.

The zeros of $NTF_a(z)$ are located in accordance with the design method of section 5.4.1. Since $NTF_w(z)$ has a DC zero, the two complex conjugate zeros of $NTF_a(z)$ are located at the same frequencies as the complex conjugate zeros of a fixed fifth order modulator (refer to table 2.1). In this way, $NTF_b(z)$ has the same zero locations as a fixed fifth order modulator and optimal noise performance will be achieved at low levels.

This system has greater design flexibility than systems A-C because the power

gain of $NTF_a(z)$ (P_n) can be adjusted by means of varying the Butterworth cutoff frequency and this allows greater control of the operating regions. The disadvantage is that the parameter set is now two-dimensional i.e. both the power gain P_n and the constant B must be chosen to optimise the dynamic range. The approach taken in designing the modulator is to obtain a set of design curves which limit the region in which B and P_n can be chosen.

Design-Space Curves

The variation in the maximum stable 1 kHz sinusoidal amplitude A_{max}, for the WBF modulator with power gain P_n is plotted in figure 5.20 for $B = 0$ and $B = \infty$. These two curves also represent the curve for fixed modulators with noise transfer functions $NTF_a(z)$ and $NTF_b(z)$ [2] respectively.

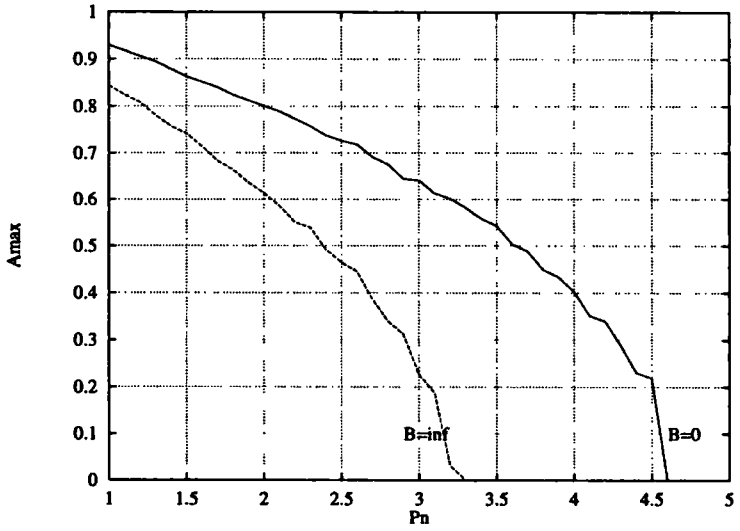


Figure 5.20: System D: A_{max} against P_n for WBF modulator with $B = 0$ and $B = \infty$.

For a point on the $A_{max}(P_n)$ curve for $B = \infty$, increasing either the amplitude of the input signal or the power gain P_n will cause the modulator to become unstable. It is possible to achieve greater stability by reducing B, allowing the NTF to adapt to the input level. The curves of $A_{max}(P_n)$ for different value of B are plotted in figure 5.21. For a given P_n, reducing B allows a greater input amplitude to be used i.e. the area between the $B = 0$ and $B = \infty$ curves represents the region where the

[2]Strictly the values of B for fixed operation is the minimum value for which the modulator enters the latent region and this is obviously smaller than $B = \infty$.

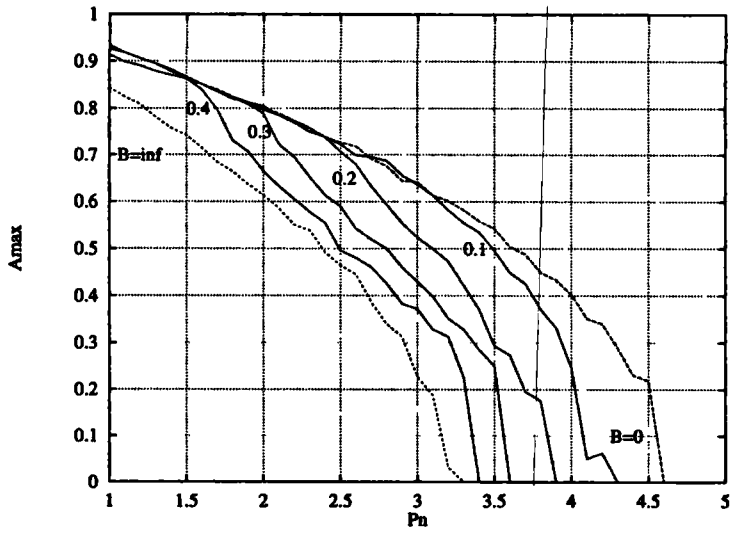


Figure 5.21: System D: A_{max} against P_n for different values of B

adaption increases the stability of the modulator. In the following this region will be referred to as the adaptive region. These curves are useful in the design process. For a specified maximum input level A_{max}, they show some of the combinations of P_n and B on the edge of the stable region. Increasing B, P_n or A_{max} from a point on the curve results in instability.

Performance Comparison with Fixed ΣΔ Modulator

We now evaluate the noise performance that can be obtained using different parameters. Additionally we consider the effects of overload on the system implementation and show how a practical implementation can restrict the level of performance obtained. It has been shown in section 5.3.3 that the overload region may be entered if the input signal A_s is too high. The depth of entry into the overload region has been evaluated in the following simulations by measuring the maximum value of $|v(k)|$ $(= ||v(k)||_\infty)$ that occurs. To relate this value to an implementation of the modulator, it has been detected when $||v(k)||_\infty$ exceeds the power-of-two overload bounds $M = 4, 8, 16$ etc. which correspond to an implementation with 2, 3, and 4 overflow bits (bits to the left of the decimal point) respectively.

Maximum Signal-to-Noise Ratio (SNR)

We first present results for the evaluation of maximum SNR that can be achieved for an input $A_s = 0.3$. The aim is to establish whether the WBF system can achieve

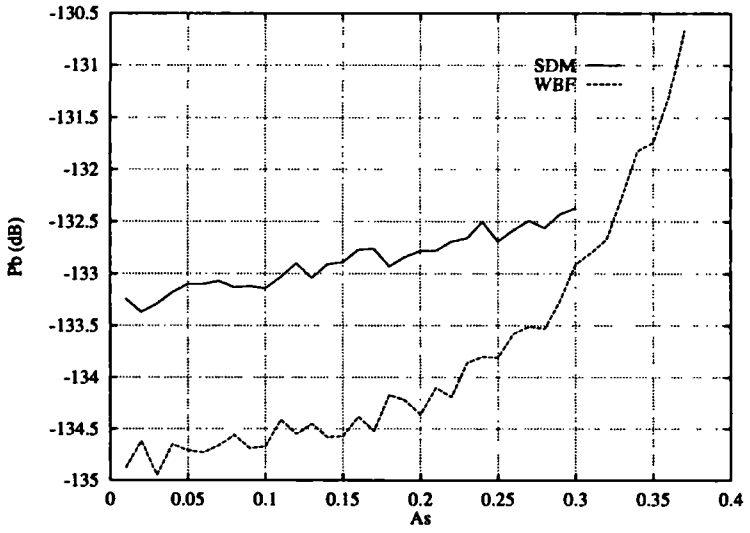


Figure 5.22: System D: Noise power against A_s for optimal WBF and fixed $\Sigma\Delta$ systems

the same performance as a fixed modulator for an input level which achieves high SNR in the fixed system. For the fixed 5th order modulator, the maximum SNR has been determined by repeated simulation with different values of power gain P_f. The NTF which produces the highest SNR has been selected. For the WBF system, the SNR has been exhaustively evaluated for different values of B and P_n. The $\{B, P_n\}$ combination which produces the highest SNR has been selected. The range of P_n to search has been established on the basis that for a fixed high order system, the maximum SNR is achieved close to the maximum input of the modulator. Hence figure 5.20 shows that for an input of $A_s = 0.3$ the search space should encompass the range $P_n = 2.8 \rightarrow 4.3$ and $B = 0 \rightarrow 1$ (simulations in appendix D.1 show that the latent region will be entered for $B > 1$ therefore the noise power will be constant with B in this region).

The results of the simulations show that the maximum SNR occurs for parameters $P_n = 2.9$, $B = 0.325$. In figure 5.22 the baseband noise power is plotted against input level for these parameters. On the same axis the results for the fixed 5th order system are plotted, with the optimal value $P_f = 4.1\ dB$. For all input levels, the WBF system achieves a lower baseband noise power, yet has a higher MSA.

The performance of the two systems is compared in table 5.4. The improvements in maximum SNR and dynamic range achieved by the WBF system over the fixed system are 0.9 dB and 3.6 dB, respectively.

	Fixed ΣΔ	WBF
Maximum SNR	121.9 dB	122.8 dB
Low-level Noise Power	-133.2 dB	-134.9 dB
Maximum input	-10.5 dB	-8.6 dB
Dynamic Range	122.7 dB	126.3 dB

Table 5.4: Performance of System D, relative to fixed fifth order modulator .

The maximum stable amplitude for the WBF modulator is $A_{max} = 0.37$. The location of $A_{max}(P_n)$ on the design space indicates that the excursion into the overload region for this input is limited since the operation is close to the edge of the adaptive region. This is confirmed by the measurement of maximum value of $|v(k)|$ obtained during the simulation: a value of $||v(k)||_\infty = 4.54$ is obtained for the input $A_s = 0.37$. The modulator can be implemented with three overflow bits.

Maximum Dynamic Range

The results for a fixed input level indicate slight improvements in SNR and dynamic range, however, they do not indicate the improvement in dynamic range that could be achieved by optimising the parameters for both maximum input level and low-level noise performance. In order to find the system parameters for maximum dynamic range, an estimate of the dynamic range has been obtained in the same way as in section 5.5.2. The maximum stable input amplitude A_{max} and the baseband noise power P_b for $A_s = 0.01$ have been found, for different values of B and P_n. For each P_n, there is a value of B which maximises the dynamic range. These values are shown in table 5.5. In figure 5.23, the dynamic range (DR) is plotted against P_n. It is crucial to note that these results are for *individually optimised systems* i.e. the system parameters are different for each point on the graph.

A maximum dynamic range of 127.1 dB is achieved for the WBF system with parameters $B = 2.7$, $P_n = 3.2$, for a maximum input $A_{max} = 0.41$. The maximum dynamic range for the fixed system is 122.7 dB with $P_f = 4.1$ dB. For the WBF system with this maximum input, the upper bound on $|v(k)|$ is $||v(k)||_\infty = 9.98$, allowing implementation with four overflow bits.

For each value of P_n there is also an associated value of A_{max} for which the maximum dynamic range is achieved. In figure 5.24, the maximum dynamic range is plotted against maximum input level, and this is shown in figure 5.24. On the

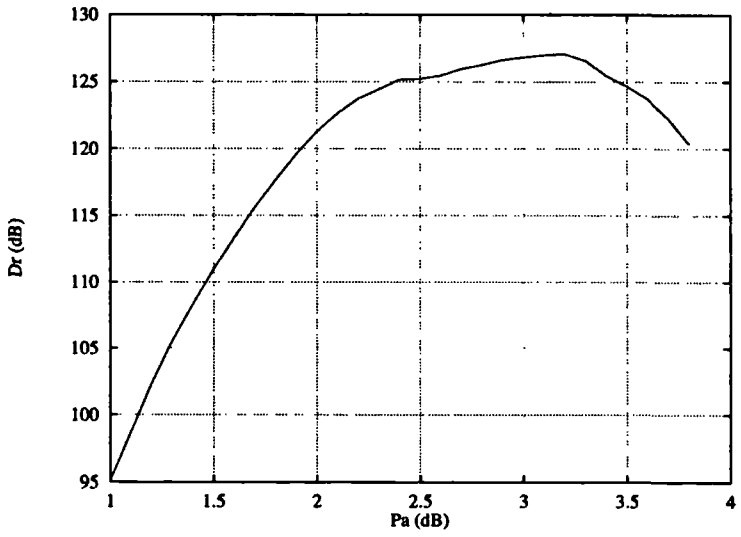


Figure 5.23: System D: Maximum Dynamic Range (DR) against power gain P_n

same axis results are plotted for a fixed 5th order system, optimised for a maximum input level A_{max} using the maximum power gain method of appendix A.4.

At high input levels, the dynamic range improvement achieved by the WBF system increases. This is because, in a fixed system the tradeoff between NTF attenuation and power gain results in high baseband noise power for systems optimised for high input levels. This tradeoff is improved in the WBF system because the NTF is adapted. The dynamic range improvement for a maximum input of $A_{max} = 0.7$ is approximately 13.3 dB. It has been shown, however, that for high A_{max}, overloading of $v(k)$ can occur. The maximum A_{max} for an overflow bound $M = 16$ (four overflow bits) is $A_{max} = 0.65$. For this value, the dynamic range improvement over the fixed modulator is 8.8 dB. In conclusion, although dynamic range improvements of over 13 dB are possible for high maximum input systems, the system implementation is limited by overload.

5.5.4 System E - Fourth Order IIR $NTF_a(z)$, Second Order FIR $NTF_w(z)$.

$NTF_a(z)$ is again a general IIR filter with Butterworth poles. $NTF_w(z)$ is an FIR filter.

Both $NTF_a(z)$ and $NTF_w(z)$ have zeros selected using the method of section 5.4.1. $NTF_b(z)$ is a sixth order filter with zeros given in table 2.1. The zeros

$P_n(dB)$	B	$P_b(dB)$	A_{max}	DR (dB)
1.0	0.24	-95.8	0.93	95.2
1.1	0.33	-99.5	0.92	98.7
1.2	0.14	-103.2	0.90	102.3
1.3	0.18	-106.5	0.90	105.6
1.4	0.20	-109.4	0.88	108.4
1.5	0.18	-112.2	0.87	111.0
1.6	0.19	-114.5	0.85	113.3
1.7	0.19	-117.2	0.84	115.6
1.8	0.21	-119.3	0.83	117.7
1.9	0.21	-121.4	0.81	119.6
2.0	0.20	-123.2	0.80	121.3
2.1	0.22	-124.8	0.79	122.6
2.2	0.24	-126.1	0.77	123.7
2.3	0.22	-126.9	0.75	124.4
2.4	0.23	-128.0	0.72	125.1
2.5	0.24	-129.0	0.65	125.2
2.6	0.27	-129.5	0.63	125.5
2.7	0.26	-130.9	0.56	125.9
2.8	0.22	-135.3	0.59	126.2
2.9	0.26	-132.7	0.50	126.6
3.0	0.24	-132.8	0.50	126.8
3.1	0.25	-133.9	0.45	127.0
3.2	0.27	-134.9	0.41	127.1
3.3	0.26	-135.1	0.34	126.6
3.4	0.24	-135.0	0.33	125.4
3.5	0.26	-135.9	0.27	124.7
3.6	0.22	-134.6	0.29	123.7
3.7	0.16	-140.8	0.34	122.2
3.8	0.15	-130.9	0.30	120.4

Table 5.5: Maximum Dynamic Range Results for System D.

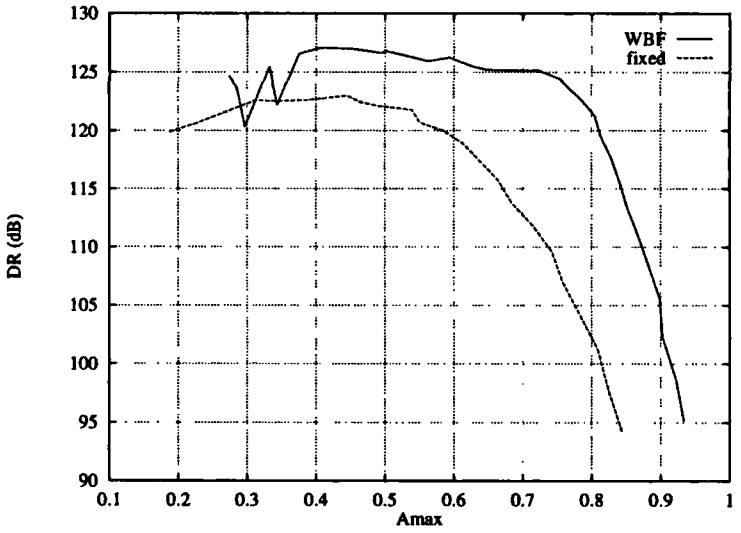


Figure 5.24: System D: Maximum dynamic range (DR) against A_{max}

of $NTF_a(z)$ are chosen from these to minimise the baseband noise power at high input levels. Expression 5.35 is minimised for complex conjugate zero frequencies (chosen from table 2.1) of ± 0.23862 and ± 0.93247, therefore these are selected for $NTF_a(z)$. The remaining complex conjugate zero location $\omega_w = \pm 0.66232$ is chosen for $NTF_w(z)$.

The dynamic range simulations of system D have been repeated for system E. The values of B for each P_n which maximise the dynamic range are given in table 5.6. In figure 5.25(a) the maximum dynamic range is plotted against A_{max}. Also plotted are the results for a fixed sixth order modulator, optimised for maximum dynamic range. For input levels above $A_{max} = 0.42$, the dynamic range of the WBF system is higher than the fixed system. The maximum dynamic range is 138.0 dB, achieved at $A_{max} = 0.72$, for parameters $P_n = 2.3$, $B = 0.26$. The improvement in dynamic range over the fixed system is 26.2 dB, a greater improvement than is achieved using system D. The 'design-space' of this modulator is plotted in figure 5.25(b). The $B = 0$ curve is the same as with system D, however the $B = \infty$ curve is not present because the modulator is unstable with $B = \infty$ for all P_n. Therefore for stability, the modulator must be adaptive over its entire operating region. Simulations have shown that due to the high gain of the weighting filter, the overload bound $M = 16$ is exceeded for all optimal system parameters. If the overload bound is increased to $M = 256$ (8 overflow bits), the modulator is implementable for $A_{max} = 0.3$, yielding $||v(k)||_\infty = 177.2$, however for this maximum input, the fixed modulator has better

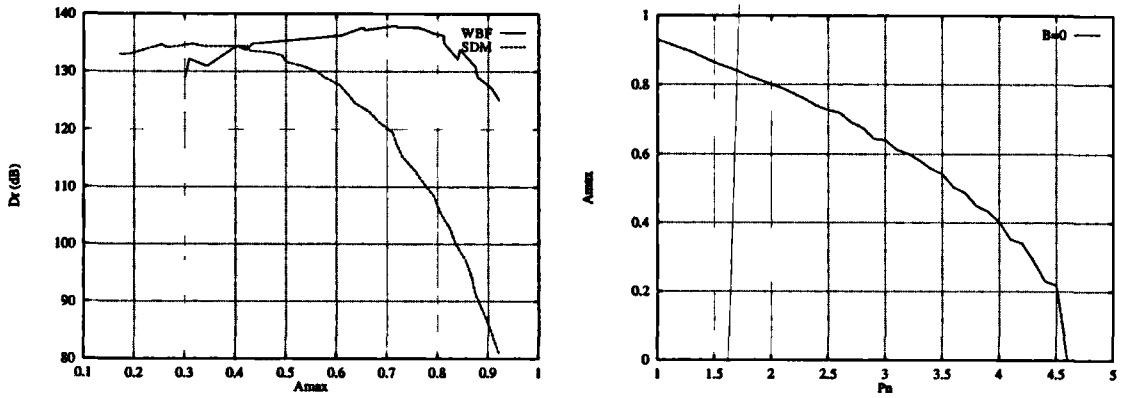


Figure 5.25: System E: (a) Maximum Dynamic Range against A_{max} (b) Design-space

performance.

In conclusion, although this system is capable of achieving substantially better dynamic range than the fixed system for high maximum inputs, severe overload prevents useful results being obtained in a real implementation.

5.6 Summary

In this chapter the technique of bit-flipping has been extended to show that it is possible to emulate a high order modulator using a lower order modulator with bit-flipping. This has been achieved by using bit-flipping to auto-correlate samples of the quantization error in such a way as to minimise the error variance, measured through a low-pass weighting filter. This causes the baseband noise to be attenuated over and above the attenuation provided by the noise shaping. The decision of whether to activate the BFO for a particular sample is governed by two conditions: the high-pass condition and the stability condition. The high-pass condition specifies which samples should be flipped to generate a high-pass error. The stability condition enhances stability by restricting bit-flipping to samples for which the quantizer input magnitude is bounded by a constant, B.

To analyse the system, an equivalent model has been derived, termed the dual quantizer model, which shows that the bit-flipping algorithm introduces local feedback around the quantizer, and enables the bit-flipping operation to be mapped onto an equivalent operation on the NTF. The model reveals that the stability condition causes the NTF to adapt according to the stability of the modulator. In this way,

$P_n(dB)$	B	$P_b(dB)$	A_{max}	DR (dB)
1.0	0.17	-125.8	0.92	125.0
1.1	0.18	-128.1	0.91	127.2
1.2	0.19	-130.2	0.88	129.0
1.3	0.20	-132.0	0.88	130.8
1.4	0.21	-133.7	0.84	132.1
1.5	0.22	-135.3	0.84	133.8
1.6	0.22	-136.8	0.81	135.0
1.7	0.23	-138.0	0.81	136.2
1.8	0.23	-138.6	0.79	136.6
1.9	0.23	-139.2	0.78	137.0
2.0	0.25	-140.0	0.76	137.7
2.1	0.25	-140.6	0.72	137.8
2.2	0.25	-140.5	0.73	137.7
2.3	0.26	-140.9	0.72	138.0
2.4	0.25	-141.4	0.65	137.6
2.5	0.27	-140.9	0.66	137.2
2.6	0.27	-140.6	0.61	136.2
2.7	0.30	-142.1	0.43	134.8
2.8	0.29	-142.3	0.40	134.4
2.9	0.28	-141.3	0.42	133.8
3.0	0.31	-142.4	0.31	132.1
3.1	0.28	-140.2	0.28	130.9
3.2	0.27	-139.5	0.27	129.0

Table 5.6: Maximum Dynamic Range Results for System E

the tradeoff between low-level noise power and maximum input level is improved.

Three operating regions have been identified which the modulator passes through during adaption: latent, transition and overload. The latent region occurs for small input signals and in this region the modulator is fixed. In the transition region, the NTF adapts rapidly with input level, causing the baseband noise power to rise. In the overload region, the local feedback loop around the quantizer opens, causing the weighting filter output to reach high signal levels. The overload region makes implementation difficult. It has been found, however, that in the case of a first order weighting filter, entry into the overload region can be restricted by increasing the power gain of the NTF. A mode of operation has also been identified where there is no latent region nor overload region, and the stability control is responsible for stability at all non-zero input levels.

The bit-flipping technique has been extended to the use of higher order weighting filters. A design technique has been proposed for the specification of NTF zero placements which maximise the dynamic range.

Some example systems have been considered and their performance compared to fixed systems. It has been shown how adaption can be used to stabilise third and fourth order Tewksbury FIR NTFs. The fourth order system achieves almost the same dynamic range as an optimal fourth order fixed modulator, without the use of NTF poles to stabilise the system. The design of a system with fourth order IIR NTF and first and second order weighting filters has been briefly investigated. It has been found that for systems optimised for high inputs, large increases in dynamic range over a fixed modulator are possible; however the degree of overload increases with maximum input level, therefore the implementation of these modulators becomes more difficult. This is especially true with the second order weighting filter, where overload is too severe to yield a practical implementation. For the first order weighting filter, a practically implementable modulator with a dynamic range increase of more than 8 dB can be obtained.

Chapter 6

Power Digital-to-Analogue Conversion using Bit-Flipping

6.1 Introduction

In chapters 4 and 5, the technique of bit-flipping has been used to enhance the linearity and dynamic range of a $\Sigma\Delta$ modulator. In this chapter, attention is turned to the use of bit-flipping to realise $\Sigma\Delta$ power D-A converters.

The technique of power digital-to-analogue conversion using a single-bit converter has been introduced in section 2.6. To recap, the basic principle is to use a 1-bit converter followed by a sample and hold to generate a two-level waveform with a low pulse-repetition frequency (PRF) and high resolution in the baseband. The waveform is amplified with a class D (switching) amplifier which regenerates the waveform at a higher voltage level. The class D amplifier comprises a power switch and a passive low-pass filter which reconstructs the baseband signal across the load (figure 6.1).

As described in section 2.6.1, previous research has largely concentrated on the

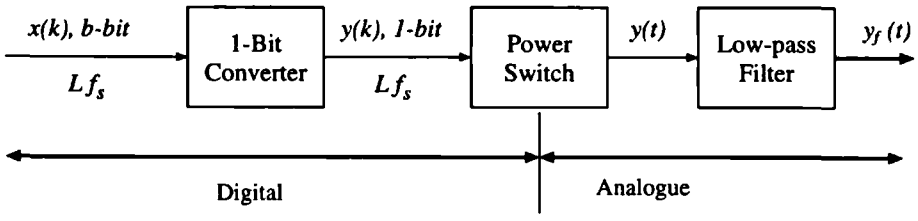


Figure 6.1: Block diagram of power D-A converter,

use of noise shaped uniformly sampled pulse-width modulation (PWM) to generate the bitstream. The main advantage of PWM is its low PRF, which translates to low power dissipation in the output stage. Noise shaped uniformly sampled PWM has three significant disadvantages, however. Firstly, it is inherently a nonlinear process, and requires a linearisation algorithm and careful NTF design if it is to be used in critical audio applications. Secondly, the clock rate to time the pulse edges is typically in the order of 90 MHz, and this rules out ASIC implementation. Thirdly, the design of the PWM stage is complicated by the requirement of guard bands to allow the output stage to recover between transitions.

The work presented in the following represents efforts to use $\Sigma\Delta$ techniques in power D-A converters. There are two prime motivating factors for doing this:

- The $\Sigma\Delta$ modulator is capable of highly linear performance, when correctly dithered (refer to chapter 4).
- The clock rate is considerably lower than PWM (typically $\approx$ 3 MHz for $L = 64$), allowing straightforward ASIC implementation and eliminating the need for guard bands.

To use a $\Sigma\Delta$ modulator for power conversion, the output samples must be first passed through a sample and hold (implemented as a clocked latch) which converts the sample impulses into a rectangular waveform of amplitude $\pm V$. The average PRF of this waveform is defined as follows, with reference to figure 6.2.

Definition 6.1 *The average pulse-repetition frequency (PRF) is defined as the reciprocal of the average time t_r between consecutive rising edges of a two-level pulse-stream.*

To date, $\Sigma\Delta$ modulators have been considered unsuitable for power conversion, due to the high rate of transitions in the 1-bit output, which result in a high PRF and high power dissipation in the power switch [Pau95]. As described in section 2.6, for real implementations of power switches, the rise and fall times are not instantaneous and so energy dissipation will occur on every pulse transition. It is therefore desirable to keep the average PRF of the bitstream as low as possible. Reports suggest a PRF around 350 kHz is suitable for power conversion up to 600 watts with efficiencies over 90% [Ped94]. The analysis presented in the following confirms that the average PRF of a $\Sigma\Delta$ bitstream exceeds this, however it will be

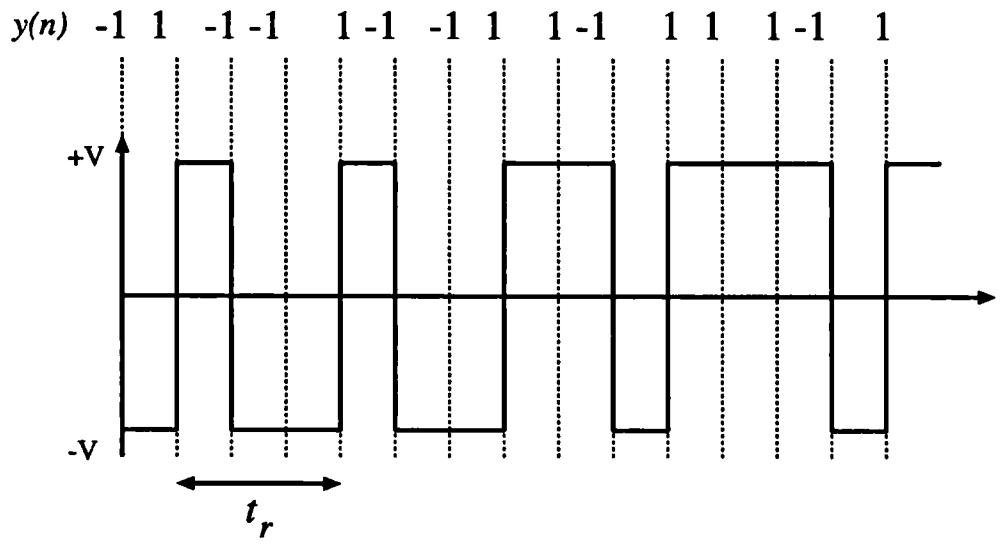


Figure 6.2: Digital signal output $y(k)$ and sample and hold output of $\Sigma\Delta$ modulator

shown that using bit-flipping it is possible to *reduce* the transition rate, making the PRF low enough for power conversion.

Initial research by the author into reducing the transition rate of the bitstream resulted in a technique termed pulse-group modulation (PGM). This technique is described in appendix E.1, which is a reprint of [Mag96a]. To summarise the technique, the modulator output samples are divided into frames and the bits in each frame are re-ordered so that all the $+1's$ and $-1's$ are grouped together. This has the effect of reducing the transitions in the bitstream. The analysis reveals that the output sequence can be modelled as noise-shaping followed by a decimation stage and PWM. Distortion and noise introduced by the pulse grouping are reduced by error-shaping around the pulse-grouping system. This system is capable of distortion and signal-to-noise ratio commensurate with 18-bit converter resolution. A disadvantage of the technique, however, is that the system complexity is high, because two noise shaping filters are required, one for the quantization error and one for the pulse-grouping error. The use of two filters makes design problematic and optimisation techniques are required to achieve high performance. Furthermore the PRF of the optimised system is twice the desirable limit for efficient power conversion.

The techniques described in this chapter utilise bit-flipping to reduce the PRF of the $\Sigma\Delta$ bitstream. This technique has the advantage that only a single noise-

shaping filter is required. It will be shown that low PRF and high resolution is possible without excessive system complexity.

Although the concern here is the development of 1-bit coding algorithms, the algorithm performance will be influenced by the characteristics of the output stage, as described in section 2.6. Therefore to properly evaluate a particular algorithm, some of the effects of non-ideal output stages must also be considered. In appendix E.2, the effect of timing errors are briefly discussed and it is shown that the PRF should remain as constant and as low as possible to increase the immunity to timing errors. For the initial results and analysis, however, it will be assumed that the output stage introduces no nonlinear distortion i.e. the bitstream is accurately represented at a high power level and perfectly reconstructed by the analogue filter.

6.1.1 Target Performance

Although the new bit-flipping algorithms will be evaluated over a range of oversampling ratios and orders, it is useful to define from the outset a target performance for the system to achieve. This will allow the results to be interpreted with regards to a practical system. For compact-disc quality digital-audio, the data storage is to 16-bits accuracy, implying a dynamic range of approximately 98 dB. This is generally regarded as the minimum acceptable performance for domestic audio systems. Recent techniques have allowed an increase of over 18 dB in the perceived dynamic range of the system by psychoacoustically optimising the shape of the noise-floor to inversely match the sensitivity of the ear [Lip91]. It is doubtful however, whether the jitter performance and accuracy of the Class D amplifier can meet the required noise performance to take advantage of this system. In this study, therefore, a target performance of 98 dB dynamic range is sought.

A target PRF 352.8 kHz is sought for the 1-bit signal, as this represents the practical upper limit for efficient power conversion [Ped94]

6.2 Pulse-Repetition Frequency of a Sigma-Delta Bitstream

In this section the PRF characteristics of a $\Sigma\Delta$ modulator are investigated. The aim is to identify the suitability of $\Sigma\Delta$ modulation for power D-A conversion. The PRF of

the bitstream has been defined in section 6.1. It can be determined from the average transition rate (ATR) of the bitstream, that is, the transitions per output sample of the signal at the input to the sample-and-hold hold. For a modulator output $y(k)$, a transition occurs before the k^{th} sample when $y(k) \neq y(k-1)$, therefore the ATR is given by:

$$\overline{T_r} = \lim_{M\to\infty} \frac{1}{2M} \sum_{k=1}^{M} |y(k) - y(k-1)| \tag{6.1}$$

The expression $|y(k) - y(k-1)|/2$ represents the number of transitions occurring at each sample. The maximum transition rate is unity, which would occur for the bitstream $\overline{1, -1}$. Here $\overline{y(k), y(k+1), \ldots, y(k+P-1)}$ represents a periodic bitstream repeating with period P. The average PRF is given by

$$\begin{aligned} \overline{f_p} &= \frac{Lf_s\overline{T_r}}{2} && (6.2) \\ &= \frac{Lf_s}{4} \lim_{M\to\infty} \frac{1}{M} \sum_{n=1}^{M} |y(k) - y(k-1)| \quad \text{(Hz)} && (6.3) \end{aligned}$$

For the *estimation* of PRF during a simulation of the modulator, the following expression is used, with the assumption made that M is large.

$$f_e = \frac{Lf_s}{4M} \sum_{n=1}^{M} |y(k) - y(k-1)| \quad \text{(Hz)} \tag{6.4}$$

6.2.1 PRF bounds with DC Input

Due to the nonlinearity of the quantizer in the loop, the $\Sigma\Delta$ modulator exhibits limit cycle phenomena, as discussed in section 2.4.1. For a rational DC input, $y(k)$ is periodic in P, i.e: $y(k) = y(k+P)$. The ATR can therefore be expressed as:

$$\overline{T_r} = \frac{1}{2P} \sum_{n=1}^{P} |y(k) - y(k-1)| \tag{6.5}$$

From equation 6.2:

$$\overline{f_p} = \frac{Lf_s}{4P} \sum_{n=1}^{P} |y(k) - y(k-1)| \tag{6.6}$$

As $y(k)$ is periodic in P, no assumptions about the size of P are required. The expression $\sum_{n=1}^{P} |y(k) - y(k-1)|$ is governed by the bit-pattern of the sequence $y(k)$.

For a given steady-state input, the possible bit patterns and the period P depend uniquely on the NTF [Ris94b]. It is possible, however, to derive a general expression for an *upper bound* on $\overline{f_p}$. The upper bound is useful because it gives the worst-case PRF, which will result in the highest power dissipation in the power switch.

Consider a sequence of period P comprising P_1 positive bits and P_2 negative bits, such that

$$P_1 + P_2 = P \tag{6.7}$$

By the DC coding property 2.1, for a rational DC input m_x:

$$\frac{P_1 - P_2}{P} = m_x \tag{6.8}$$

Combining equations 6.7 and 6.8:

$$\frac{P_1}{P} = \frac{1 + m_x}{2} \tag{6.9}$$

The ratio P_1/P can thus be evaluated and the minimum period P is obtained by reducing P_1/P to its lowest possible terms. In table 6.1 some of the possible bit patterns which satisfy equation 6.9 with minimum period P are presented for different positive DC inputs m_x. For these examples the highest transition rate occurs when each -1 sample is preceded and followed by a sample of value $+1$, therefore there are *two transitions* associated with every -1. These cases are shown in bold type in the table. Similarly, where $m_x < 0$ (not shown) the highest transition rate occurs when each $+1$ sample is preceded and followed by a sample of value -1. The maximum transition rate for $m_x \geq 0$ is therefore $\hat{T}_r = 2P_2/P$ and the maximum transition rate for $m_x < 0$ is $\hat{T}_r = 2P_1/P$.

- For $m_x \geq 0$

  From equation 6.7:

$$\hat{T}_r = \frac{2(P - P_1)}{P} \tag{6.10}$$

  From equation 6.8

$$\begin{aligned} \hat{T}_r &= 2(1 - \frac{1 + m_x}{2}) \\ &= 1 - m_x \end{aligned} \tag{6.11}$$

- For $m_x < 0$

$$\hat{T}_r = \frac{2P_1}{P} = 1 + m_x$$

m_x	P_1/P	Example Minimum Period Sequences	$\overline{T_r}$
0	1/2	+1 -1	**1**
1/2	3/4	+1 +1 -1 +1	**1/2**
1/3	2/3	+1 -1 +1	**2/3**
2/3	5/6	+1 +1 +1 -1 +1 +1	**1/3**
1/4	5/8	+1 +1 +1 +1 -1 -1 -1 +1	1/4
		+1 +1 -1 +1 -1 -1 +1 +1	1/2
		+1 -1 +1 -1 +1 -1 +1 +1	**3/4**
3/4	7/8	+1 +1 +1 +1 -1 +1 +1 +1	**1/4**
1/5	3/5	+1 +1 -1 -1 +1	2/5
		+1 -1 +1 -1 +1	**4/5**
2/5	7/10	+1 +1 +1 +1 -1 -1 -1 +1 +1 +1	1/5
		+1 +1 +1 -1 +1 -1 -1 +1 +1 +1	2/5
		+1 +1 +1 -1 +1 -1 +1 -1 +1 +1	**3/5**
3/5	4/5	+1 +1 +1 -1 +1	**2/5**
4/5	9/10	+1 +1 +1 -1 +1 +1 +1 +1 +1 +1	**1/5**

Table 6.1: Example minimum-period sequences and transition rate for DC input m_x

These expressions can be combined to give

$$\hat{T_r} = 1 - |m_x| \tag{6.12}$$

leading to an expression for the maximum PRF with DC input m_x:

$$\hat{f_{pd}}(m_x) = \frac{Lf_s}{2}(1 - |m_x|) \tag{6.13}$$

The PRF reaches a theoretical maximum of $LF_s/2$ for $m_x = 0$, corresponding to a periodic bit pattern $\overline{1,-1}$. For a positive input, the PRF falls linearly as m_x increases and the ratio P_1/P increases (by equation 6.9), implying that a greater percentage of samples must have the value +1 (i.e. the pulse 'density' increases). The number of occurrences of -1 reduces and the transition rate is reduced. Similarly, with a negative DC input, the ratio P_2/P increases and the number of occurrences of +1 reduces. The PRF reaches zero for $|m_x| = 1$, corresponding to a bit pattern comprising of $y(k) = +1$ for all k, or $y(k) = -1$ for all k. Note that in the above expressions, there is no inclusion of the NTF and the only condition for the PRF to be bounded by 6.13 is that the DC coding constraint is satisfied. This constraint implies firstly that the modulator has high noise attenuation at DC, and secondly, that the modulator is stable.

6.2.2 PRF bounds with Sinusoidal Input

An expression for the maximum PRF for a low-frequency sinewave can be derived by interpreting the sinewave input as a slowly varying DC input. This interpretation has been discussed in section 2.2.3. The maximum PRF is found by averaging the PRF over all the instantaneous amplitudes of the sine function. Due to waveform symmetry, the PRF can be evaluated over a quarter waveform:

$$\hat{f_{ps}}(A_s) = \frac{1}{\pi/2}\int_0^{\pi/2} \hat{f_{pd}}(A_s \sin\alpha)d\alpha \tag{6.14}$$

where A_s is the peak sine amplitude and $\hat{f_{pd}}(m_x)$ is the maximum PRF for DC input m_x.

From equation 6.13:

$$\hat{f_{ps}}(A_s) = \frac{Lf_s}{\pi}\int_0^{\pi/2}(1 - |A_s \sin\alpha|)d\alpha \tag{6.15}$$

$$= \frac{Lf_s}{2}\left(1 - \frac{2A_s}{\pi}\right) \qquad \text{where } A_s > 0 \tag{6.16}$$

Again, a maximum PRF of $Lf_s/2$ is obtained for zero input. The PRF reduces linearly with A_s. For sinewave amplitudes $A_s > 0$ the PRF is greater than with the equivalent DC input $m_x = A_s$. This is because the PRF integral has contributions from levels smaller than m_x which are associated with instantaneously higher PRFs than $\hat{f_{pd}}(m_x)$.

6.2.3 Accuracy of Bounds

In this section, results are provided to verify the theoretical upper bound PRF approximations of section 6.2. The results concentrate on a modulator with parameters $L = 64$, $N = 4$ and power gains in the range $P_n = 1 \rightarrow 4\ dB$.

As described above, the upper bounds for DC and sinusoidal inputs are denoted by $\hat{f_{pd}}$ and $\hat{f_{ps}}$. The corresponding experimental *average* PRFs are given by $\overline{f_{pde}}$ and $\overline{f_{pse}}$, measured in a simulation of 100000 samples.

The PRFs against DC and 1 kHz sinusoidal input amplitudes are plotted in figures 6.3 and 6.4. On the same axis, the values of $\hat{f_{pd}}$ and $\hat{f_{ps}}$ are plotted. The maximum theoretical PRFs successfully bound the experimental values. The values of $\overline{f_{pde}}$ and $\overline{f_{pse}}$ converge on $\hat{f_{pd}}$ and $\hat{f_{ps}}$ as the input level increases. This convergence is especially good with the DC input signal and for low power gain filters. This shows that the DC model is highly accurate under these conditions. The sinusoidal model also has good accuracy though the upper bound is not as tight.

The simulations have been repeated with a 20 kHz input signal, and the results are shown in figure 6.5. The PRF is oscillatory in nature and the upper bound is exceeded slightly at the peaks of the oscillation. A possible cause of error in $\hat{f_{ps}}$ is that the assumption that the sinewave can be represented as a slowly varying DC input is less accurate for higher frequency inputs. The reason for the oscillatory nature of the measured PRF is that the modulator has strong limit cycles which are excited by the high frequency input in a manner which is non-uniform with input level.

The convergence on $\hat{f_{pd}}$ and $\hat{f_{ps}}$ for higher input levels confirms that the output sample values are well separated i.e. for positive inputs, the majority of samples of value -1 are preceded and followed by $+1$, ensuring that the transition rate is high. An explanation for this is that at high positive input levels, the ratio P_1/P is high, therefore the density of -1 samples is low and the likelihood of adjacent -1 samples occurring is low. At lower input levels, and especially for higher power

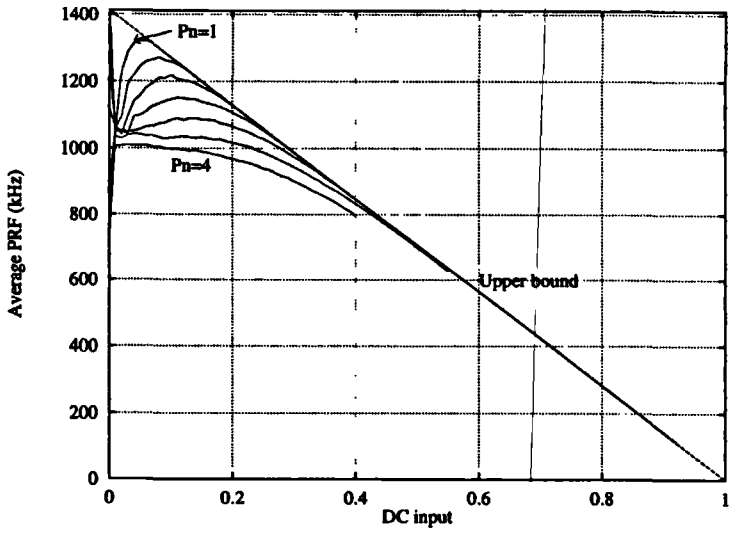


Figure 6.3: Average PRF against input level for $L = 64$, $N = 4$ modulators with NTF power gains $P_n = 1\ dB \rightarrow 4\ dB$ for DC input. Also shown is PRF bound $\hat{f_{pd}}$ (upper curve).

gains, the measured PRF is considerably lower than the maximum, indicating that groups of adjacent -1 samples occur more frequently. This is because there is an increased 'density' of -1 samples as the ratio P_1/P becomes closer to unity, and the probability that a -1 will precede or follow another -1 is greater.

The behaviour of the modulator at zero input is especially interesting, with the PRF exhibiting a 'jump' at zero input. This occurs because, for zero input, the modulator output becomes dominated by strong periodic limit cycles, as characterised by a tonal quantization error (refer to chapter 4).

6.3 Bit-Flipping Algorithms for Sigma-Delta Power D-A conversion

The analysis and results indicate that the PRF is signal dependent, in terms of the signal amplitude, frequency and type. It is shown in appendix E.2 that in the context of a sigma-delta power D-A converter, this signal dependency causes increased sensitivity to non-idealities in the power switching stage, resulting in poor distortion and noise performance, when compared to PWM-based converters. A further disadvantage of $\Sigma\Delta$ modulation, indicated by the results, is that the maximum PRF is

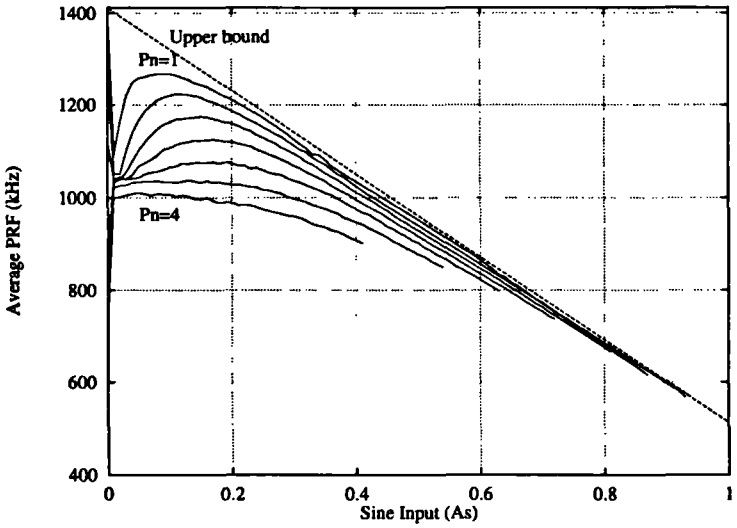


Figure 6.4: Average PRF against input level (1 kHz sinewave) for $L = 64$, $N = 4$ modulators with NTF power gains $P_n = 1\ dB \rightarrow 4\ dB$. Also shown is PRF bound $\hat{f_{ps}}$ (upper curve)

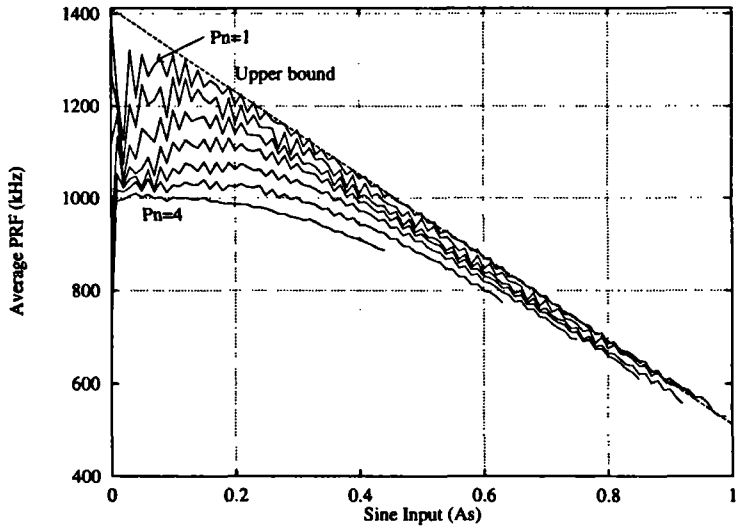


Figure 6.5: Average PRF against sinewave input level (20 kHz sinewave) for $L = 64$, $N = 4$ modulators with NTF power gains $P_n = 1\ dB \rightarrow 4\ dB$. Also shown is maximum PRF bound $\hat{f_{ps}}$ (upper curve)

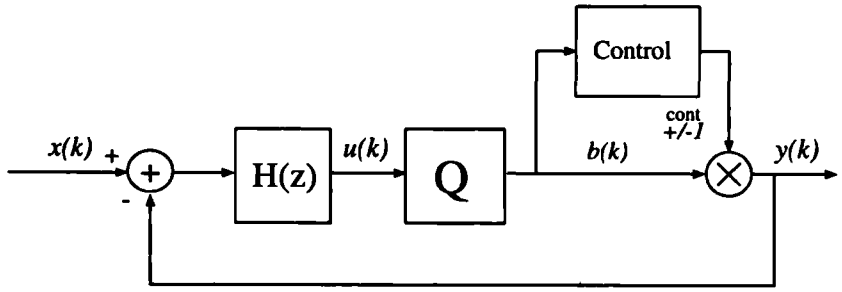


Figure 6.6: General block diagram of ΣΔ modulator with bit-flipping Algorithm

well in excess of the nominal 352.5 kHz proposed by other researchers. Using a lower oversampling ratio, the order of the modulator must be increased. For $L = 32$, a seventh order modulator is required to achieve 98 dB dynamic range (refer to section 6.6.2).

These factors have led previous researchers to the conclusion the the ΣΔ modulator is unsuitable for power converters. In the remainder of this chapter, it will be shown that by using bit-flipping, these objections can be overcome and significant advantages can be gained over noise-shaped PWM.

The basis of the bit-flipping algorithm is to invert the state of selected samples in order to reduce the PRF. The general topology of the modulator is shown in figure 6.6. An example of the concept of PRF reduction using bit-flipping is shown in figure 6.7. The lower sequence has a lower PRF than the upper sequence due to the 'reduction' of the high transition rate {1, -1} pattern. By *reduction* we mean the replacement of a high transition rate bit pattern by a lower frequency pattern. In this section the design of an algorithm to control the bit-flipping is discussed. As with the other bit-flipping algorithms proposed in this dissertation, the bit-flipping operator (BFO) is within the feedback loop, to ensure that the bit-flipping error is post-compensated by the quantizer decision in future samples.

6.3.1 PRF Control

Investigations begin with an algorithm which controls the bit-flipping rate, with the aim of making the PRF constant and signal-independent. The algorithm operates by detecting transitions in the bitstream and using bit-flipping to selectively reduce the transitions. The PRF control operates by only allowing the BFO to operate if an estimate of the average PRF exceeds a specified constant.

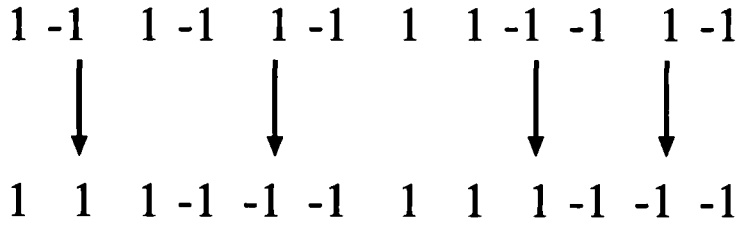


Figure 6.7: Transition reduction using bit-flipping. The upper sequence is bit-flipped to produce the lower sequence

The transitions are detected by comparing the current and previous quantizer output. A transition occurs if $y(k) \neq y(k-1)$. An estimate of the PRF is obtained by evaluating equation 6.4 up to the current sample $k = M$:

$$f_e(k) = \frac{Lf_sT(k)}{2k} \tag{6.17}$$

where k is the current sample number and $T(k)$ is the number of transitions in the previous k samples i.e.:

$$T(k) = \frac{1}{2}\sum_{i=0}^{k} |y(i) - y(i-1)| \tag{6.18}$$

Defining the system constant, f_t, as the target maximum PRF of the bitstream, the PRF bit-flipping condition is to invert $y(k)$ for samples in which the following two conditions are met:

$$y(k) \neq y(k-1) \tag{6.19}$$

$$f_e(k) > f_t \tag{6.20}$$

Notice that the bit-flipping only attempts to reduce the average PRF if it is too high, but no attempt is made to increase the PRF if it is too low. This will be justified in the next section, where it will be shown that the occurrence of $f_t > f_e(k)$ is very unlikely in a practical system.

The algorithm can be simplified by combining condition 6.20 with equation 6.17:

$$\frac{T(k)}{k} > \frac{2f_t}{Lf_s} \tag{6.21}$$

The PRF constant is now defined:

$$F_k = Lf_s/2f_t \tag{6.22}$$

The use of a PRF constant simplifies the PRF control algorithm. Condition 6.21 can be now re-expressed as:

$$\frac{T(k)}{k} > \frac{1}{F_k} \tag{6.23}$$

$$F_kT(k) > k \tag{6.24}$$

$$F_kT(k) - k > 0 \tag{6.25}$$

This can be implemented as a counter which adds F_k for every transition and counts down one on every sample. The PRF condition is TRUE if the counter output is positive. For $F_k = 1$, $k \geq T(k)$ for all k, therefore the bit-flipping algorithm becomes inactive and the operation of the modulator is that of an unmodified modulator. The flow chart of the algorithm is shown in figure 6.8.

Constant PRF Bounds

In this section the expressions for maximum PRF are used to derive expressions for a condition in which the bit-flipping algorithm keeps the PRF constant. For this to happen, the average PRF of the modulator with bit-flipping must be greater than the PRF without bit-flipping. It has been shown in section 6.2 that the latter is dependent on the signal type as well as amplitude, therefore conditions for constant PRF are given for the cases of DC and sinusoidal inputs.

- DC inputs:

  Due to the constraint imposed by the PRF condition, the conditions for the algorithm to be rendered inactive for a DC input is $f_t > \hat{f_{pd}}(m_x)$.

  From equation 6.13:

$$f_t > \frac{Lf_s}{2}(1 - |m_x|) \tag{6.26}$$

  From equation 6.22

$$|m_x| > \frac{F_k - 1}{F_k} \tag{6.27}$$

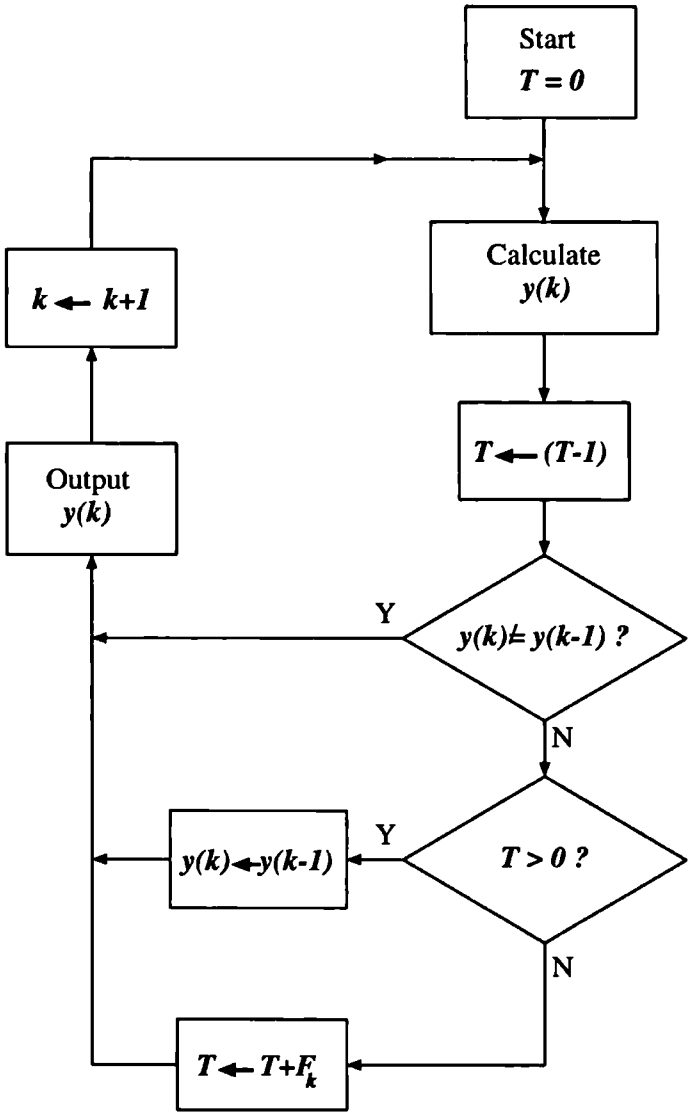


Figure 6.8: Flow chart of bit-flipping algorithm with PRF constraint

F_k	$\lvert m_x \rvert$	A_s
1	0	0
2	1/2	0.785
3	2/3	1.047
4	3/4	1.178
5	4/5	1.257

Table 6.2: Upper bound on $|m_x|$ and A_s for constant PRF

- Sinusoidal inputs:

  From equation 6.16:

$$f_t > \frac{Lf_s}{2}\left(1 - \frac{2A}{\pi}\right) \tag{6.28}$$

  leading to

$$A > \frac{\pi(F_k - 1)}{2F_k} \tag{6.29}$$

When conditions 6.27 or 6.29 are satisfied for DC and sinusoidal inputs, respectively, the bit-flipping algorithm turns off and the PRF no longer remains constant. The values of $|m_x|$ and A_s for $F_k = 1 \rightarrow 5$ are shown in table 6.2. For DC inputs, the maximum $|m_x|$ for constant PRF increases with F_k. For sinusoidal inputs, and $F_k \geq 3$ the maximum A_s exceeds full scale ($A_s = 1$) therefore constant PRF is achieved at all input levels for $F_k \geq 3$. Note that these values represent *upper* bounds for $|m_x|$ and A_s: the assumption is made that the PRF of the modulator is maximal. It is also important to note that for practical modulator implementations, optimal noise performance is usually achieved (for sinewave inputs) at input levels around $A_s = 0.25 \rightarrow 0.35$ [Ris94b]. Therefore, for practical situations, constant PRF will be obtained for all $F_k > 1$.

Examples

Simulations of average PRF against DC and sinewave input (section 6.2.3) have been repeated with the bit-flipping algorithm for $F_k = 1 \rightarrow 5$ (figures 6.9 and 6.10) and the modulator {64, 4, 2.0}. With $F_k = 1$ the modulator is identical to a standard $\Sigma\Delta$ modulator. With $F_k = 2$ the bit-flipping is active and the PRF

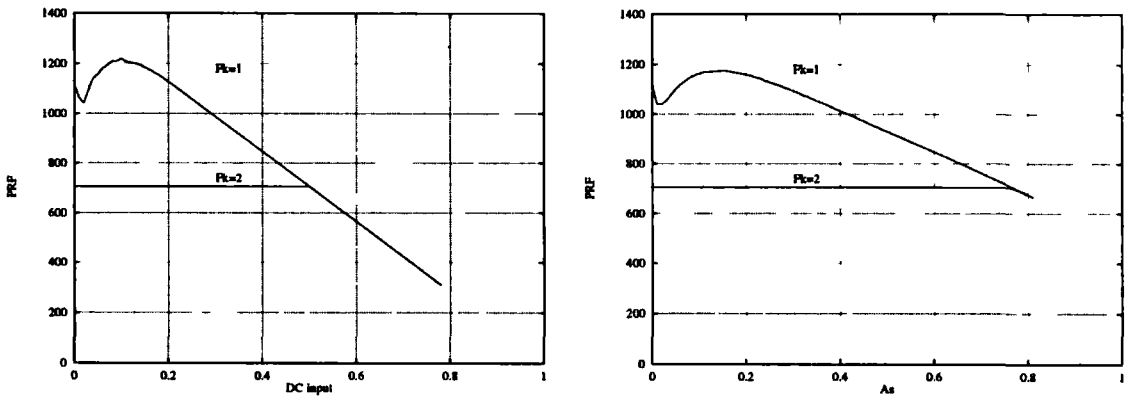


Figure 6.9: Average PRF against input level for bit-flipping modulator {64, 4, 2.0} with PRF control and $F_k = 1, 2$ (a) DC, (b) 1 kHz sinewave.

remains a constant 705.6 kHz (as predicted by equation 6.22) until the PRF of the unmodified modulator falls below the PRF of the bit-flipping algorithm i.e. until the conditions represented by equations 6.27 or 6.29 are satisfied. The PRF is also independent of signal type.

The bit-flipping reduces the stability of the modulator and for $F_k > 2$ the system is unstable at all input levels. The noise power for $F_k = 1$ and $F_k = 2$ is -97.4 dB and -90.6 dB, respectively i.e. the bit-flipping increases the noise power. The stability and noise performance will be investigated in further detail in section 6.6.3 for modulators with different power gains. In figure 6.11 the wideband spectrum is shown for the above parameters with $F_k = 2$ and a 1 kHz sinewave input:$A_s = 0.3$. A high amplitude tone occurs at the PRF of 705.6 kHz due to periodicity in the bitstream.

6.3.2 Improving Stability Margins and Noise Performance

The stability of $\Sigma\Delta$ modulators utilising bit-flipping has been investigated in chapters 4 and 5. It has been shown that bit-flipping increases the magnitude of the instantaneous quantization error, causing the stability to be compromised and the baseband noise power to increase. In this section modifications to the algorithm are considered which enhance the stability of the modulator. Two techniques are investigated, used in conjunction with the PRF condition.

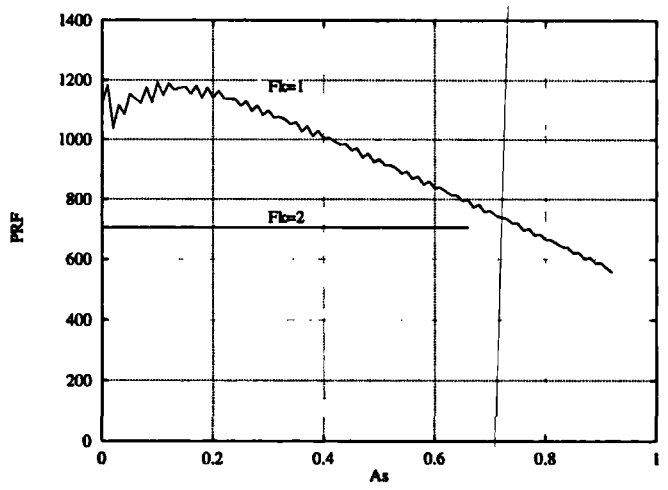


Figure 6.10: Average PRF against input level (20 kHz sinewave) for bit-flipping modulator {64, 4, 2.0} with PRF control and $F_k = 1, 2$.

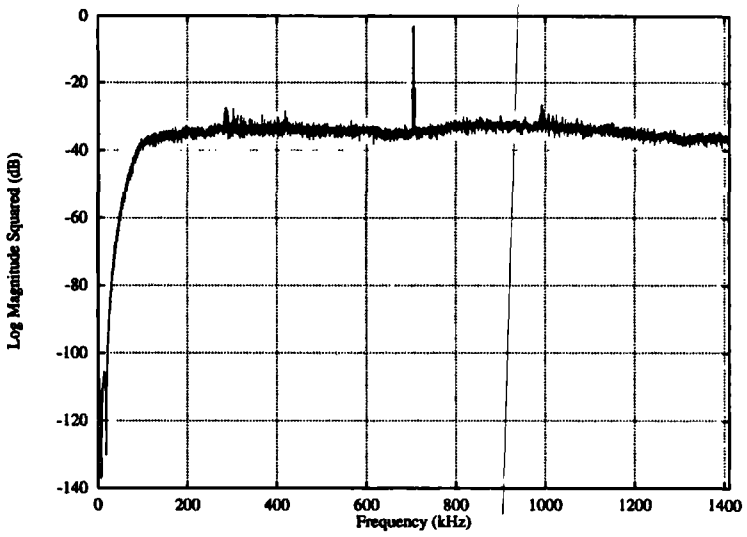


Figure 6.11: Wideband spectrum of bit-flipping system output with PRF constraint and $F_k = 2$

Quantizer Input Bound

A technique has been proposed in chapter 5 to reduce the quantizer error magnitude of flipped samples by only flipping the quantizer state if the quantizer input magnitude $|u(k)|$ is bounded by a constant chosen to yield adequate modulator stability. This constant B is called the quantizer input bound .

The algorithm operation using the quantizer input bound is to invert the quantizer output for all samples in which the following three conditions are satisfied:

$$y(k) \neq y(k-1) \tag{6.30}$$

$$f_e(k) > f_t \tag{6.31}$$

$$|u(k)| < B \tag{6.32}$$

The flow chart of the algorithm is shown in figure 6.12

Alternation Constraint

We also investigate a second technique termed the alternation constraint which aims to shape the error introduced by the BFO. Unlike the WBF algorithms of chapter 5, only the bit-flipping error is shaped, rather than the combined quantizer and bit-flipping error. The baseband quantization noise therefore still increases due to the bit-flipping, but to a lesser extent that with the PRF condition alone.

We begin by defining the error sequence of the BFO:

$$e_b(k) = y(k) - b(k) \tag{6.33}$$

where $b(k)$ is the output of the quantizer and $y(k)$ is the output of the BFO.

$e_b(k)$ has a value of $+2$ for every positive bit-flip and -2 for every negative bit-flip. An approximation of the low frequency content of $e_b(k)$ can be found by passing the error through a discrete-time integrator, which emphasises low frequencies:

$$e_w(k) = \sum_{i=1}^{k} e_b(i) \tag{6.34}$$

A method of ensuring that the peak-peak value of $e_w(k)$ is minimised is to minimise the number of equi-sign flips. The tightest constraint is that every $1 \rightarrow -1$

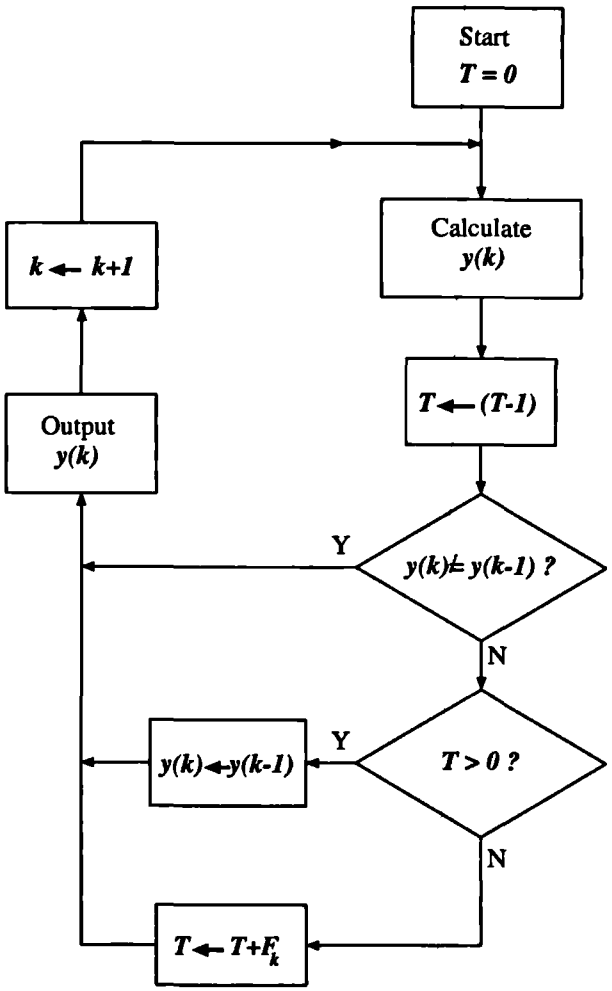


Figure 6.12: Flow chart of bit-flipping algorithm with quantizer input bound and PRF constraint

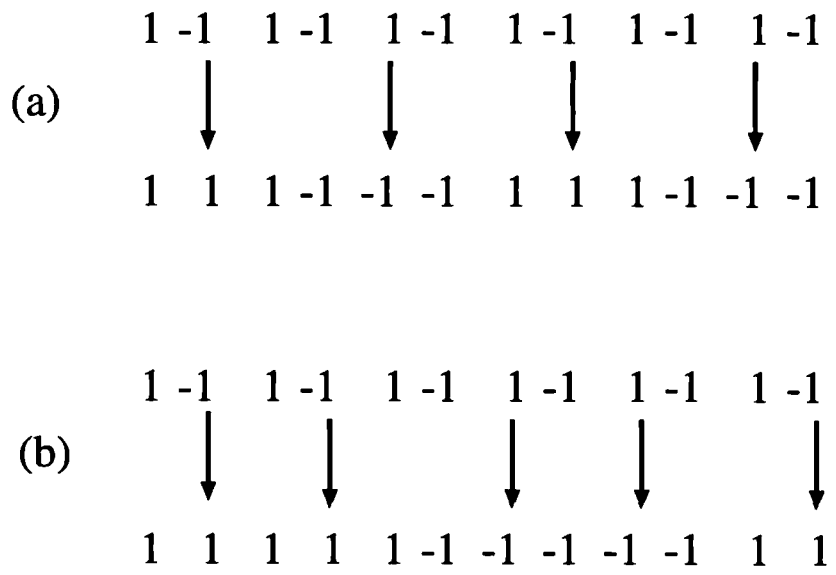


Figure 6.13: Bit patterns at input and output of bit-flipping operator with a) $S_a = 1$ and b) $S_a = 2$

flip must be followed by a $-1 \rightarrow 1$ flip and vice-versa. This ensures that the maximum peak-peak value of $e_w(k)$ is minimised to 2. The constraint may be relaxed by increasing the number of allowable consecutive equi-sign bit-flips. Examples of this are given in figure 6.13. Counters are used to measure the peak positive and negative value of $e_w(k)$ by counting the number of positive (i.e. $-1 \rightarrow +1$) and negative (i.e. $+1 \rightarrow -1$) bit flips. Details of the algorithm used are given in figure 6.14. The constant S_a is termed the alternation factor and defines the maximum counter value allowable for bit-flipping.

In table 6.3, $e_b(k)$ and $e_w(k)$ have been evaluated for an example bit-flipping sequence with $S_a = 2$. The two counter values for positive and negative bit-flips are also shown (C_p and C_m). Note that in this example only the samples in which a bit-flip occurs are shown. It can be seen that the peak-peak value of the weighted error $e_w(k)$ is limited to $2S_a$. Limiting the peak-peak value of $e_w(k)$ also has the effect of limiting the variance of the weighted error, by reducing the maximum sample-by-sample variance contribution to the value: S_a^2. The effect is to control the baseband noise power introduced by the bit-flipping.

The operation of the bit-flipping algorithm with the alternation constraint active is now defined more precisely: invert the quantizer output for all samples in which the following conditions apply:

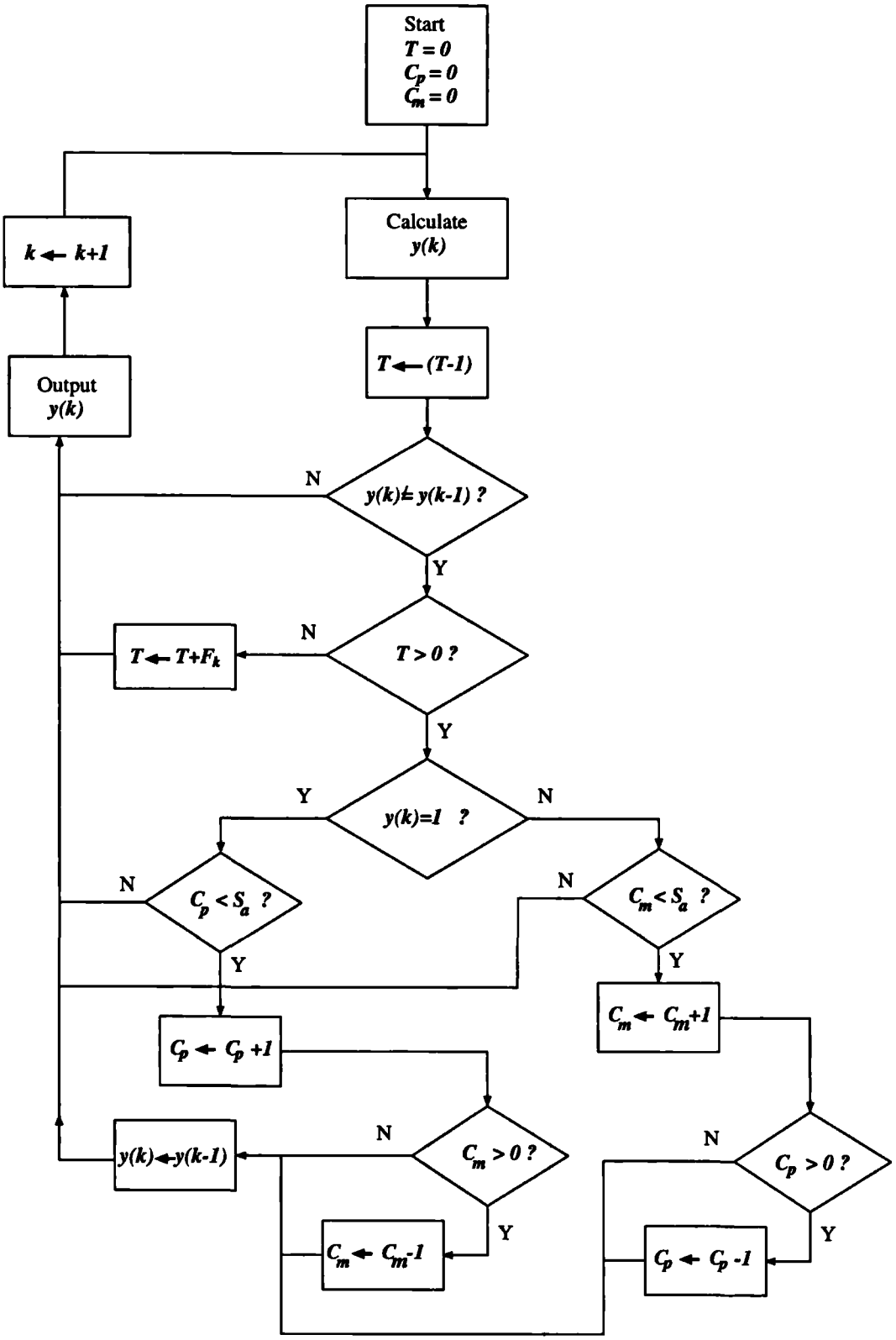


Figure 6.14: Flow chart of bit-flipping algorithm with alternation constraint and PRF constraint

bit-flipping	C_m	C_p	$e_b(k)$	$e_w(k)$
$+1 \rightarrow -1$	0	1	-2	-2
$-1 \rightarrow +1$	1	0	2	0
$-1 \rightarrow +1$	2	0	2	2
$+1 \rightarrow -1$	1	1	-2	0
$+1 \rightarrow -1$	0	2	-2	-2
$-1 \rightarrow +1$	1	1	2	0
$+1 \rightarrow -1$	0	2	-2	-2
$-1 \rightarrow +1$	1	1	2	0
$-1 \rightarrow +1$	2	0	2	2

Table 6.3: Values of alternation counters and peak bit-flipping error for $S_a = 2$

$$y(k) \neq y(k-1) \tag{6.35}$$

$$f_e(k) > f_t \tag{6.36}$$

$$|e_{wa}(k)| < S_a \tag{6.37}$$

where $e_{wa}(k)$ is the AC component of $e_w(k)$.

Examples

To demonstrate the improvements in stability possible using the quantizer input bound and alternation constraint, simulations have been performed to evaluate the PRF versus input amplitude for a 1 kHz sinewave input using the modulator {64, 4, 2.0}.

We first consider the quantizer input bound. In figures 6.15 and 6.16 the PRF is plotted against a sinewave input of frequency 1 kHz, for values of $B = 0.1$ and $B = 0.3$.

With $B = 0.1$, for both filters, the condition $|u(k)| < B$ places a lower bound on the PRF, effective for all values of F_k. The possible reduction in PRF is severely limited because the condition $|u(k)| < B$ is not satisfied for sufficient samples to ensure the PRF is kept constant by the bit-flipping. This limitation is especially a problem at low input levels, where the PRF of the standard modulator is very high and more bit-flipping is required to reduce it. At higher input levels, the variance of $u(k)$ increases and therefore the bit-flipping condition is satisfied for fewer samples.

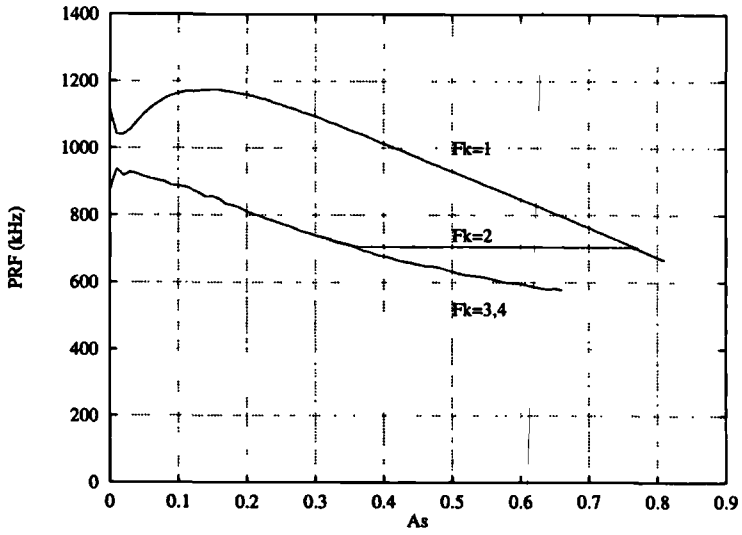


Figure 6.15: Average PRF against input level (1 kHz sinewave) for bit-flipping modulator {64, 4, 2.0} with PRF control and quantizer input bound: $B = 0.1$

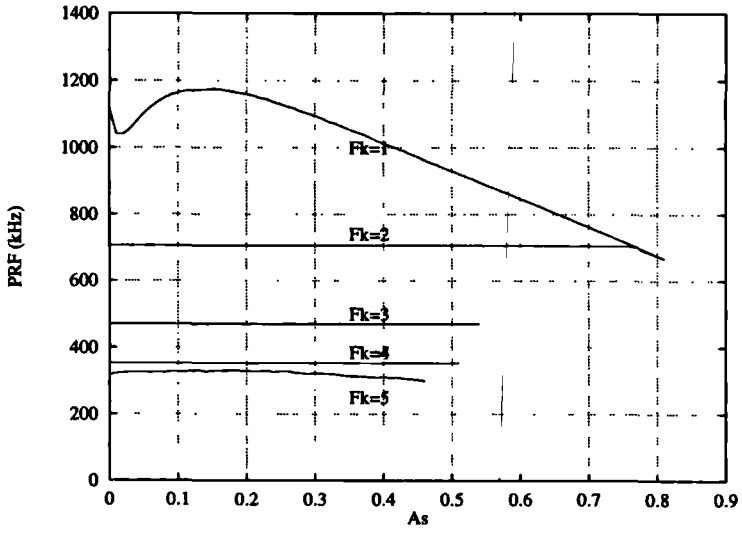


Figure 6.16: Average PRF against input level (1 kHz sinewave) for bit-flipping modulator {64, 4, 2.0} with PRF control and quantizer input bound: $B = 0.3$

F_k	$B = 0.1$			$B = 0.3$		
	PRF (kHz)	P_b (dB) ($A_s = .01$)	MSA	PRF (kHz)	P_b (dB) ($A_s = .01$)	MSA
1	665.9 → 1173.8	-97.4	0.81	665.9 → 1173.8	-97.4	0.81
2	674.0 → 937.5	-88.4	0.8	674.0 → 705.6	-86.2	0.8
3	577.5 → 937.5	-88.4	0.66	470.4	-88.5	0.54
4	577.5 → 937.5	-88.4	0.66	352.8	-89.0	0.51
5	577.5 → 937.5	-88.4	0.66	299.6 → 332.1	-89.5	0.46

Table 6.4: Simulation results for bit-flipping modulator {64, 4, 2.0} with PRF control and quantizer input bound

This limits the reduction in PRF possible, therefore the difference between the lower PRF bound and the upper PRF bound (i.e. the PRF of the unmodified modulator) is lower. With $B = 0.3$, the restriction is lifted, with the exception of $F_k = 5$, because sufficient bit-flipping is now possible for the PRF to remain essentially constant. For the higher power gain NTF, the quantizer bound more seriously limits the PRF because there is an amplification in the power of the circulating quantization noise which causes the variance of $u(k)$ to become larger.

The results are summarised in table 6.4. Generally, the stability of the modulator deteriorates with F_k, observed by a reduction in maximum stable input amplitude. For $B = 0.1$ the noise power, measured at an input of $A_s = 0.01$ remains constant for $F_k > 1$. This is because the lower bound imposed by the quantizer input bound at this level makes the PRF and consequently the flipping rate constant with F_k. With $B = 0.3$ the noise power now increases with F_k because the lower bound on the PRF is relaxed.

We now consider the alternation constraint. In figures 6.17 and 6.18 results are shown for $S_a = 1$ and $S_a = 2$.

The alternation constraint places a similar lower bound on the PRF as the quantizer input bound. The lower bound imposed by the alternation constraint is less restrictive than the one imposed by the quantizer input bound, though it should be noted that this is only specific to the constants S_a and B chosen here.

Focussing on stability and noise aspects (table 6.5), there is a general reduction in stability with F_k and this is radically observed for $S_a = 2$. For $S_a = 1$, the noise power remains fairly constant as F_k increases. For $F_k = 4$ and $F_k = 5$, the

F_k	$S_a = 1$			$S_a = 2$		
	PRF (kHz)	P_b (dB) ($A_s = .01$)	MSA	PRF (kHz)	P_b (dB) ($A_s = .01$)	MSA
1	665.9 → 1174.6	-97.4	0.81	665.9 → 1174.6	-97.4	0.81
2	673.8 → 705.6	-96.4	0.80	705.6	-88.5	0.06
3	470.4	-96.9	0.42	-	-	-
4	352.8	-97.8	0.43	-	-	-
5	286.1 → 347.9	-97.9	0.40	-	-	-

Table 6.5: Simulation Results for bit-flipping modulator {64, 4, 2.0} with PRF control and alternation constraint

noise power actually *reduces* slightly. This is due to a dithering action obtained by the bit-flipping (refer to chapter 4) causing the attenuation dominant limit cycle oscillations in the modulator which cause baseband tones.

Finally, in figure 6.19 the wideband spectrum of the bit-flipping system with PRF and alternation constraint is plotted, for parameters {64, 4, 1.0}, $F_k = 5$, $S_a = 1$, $A_s = 0.3$. A spectral peak in the noise floor occurs at approximately 395 kHz, corresponding to the average PRF. This will be investigated more closely in the next section.

6.4 Analysis of Bit-Flipping with Alternation Constraint

In this section a particular case of the bit-flipping algorithm is analysed, in which the alternation algorithm dominates the PRF control. This case occurs in the example of figure 6.19 for $F_k = 5$, where the minimum possible PRF is restricted by the alternation constraint. Assuming that the PRF control does not influence the bit-flipping, the algorithm is greatly simplified: invert $y(k)$ for all samples which satisfy

$$y(k) \neq y(k-1) \tag{6.38}$$

$$|e_{wa}(k)| < S_a \tag{6.39}$$

We being the analysis of this algorithm with the definition of a parameter N_{min}:

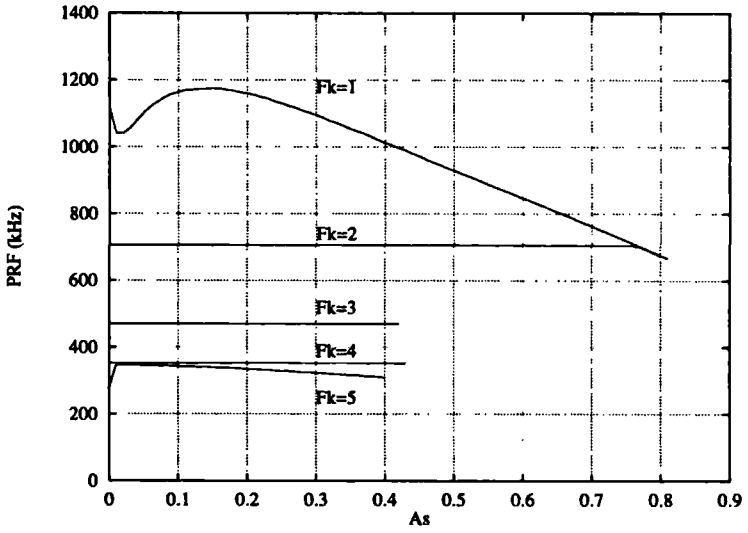


Figure 6.17: Average PRF against input level (1 kHz sinewave) for bit-flipping modulator {64, 4, 2.0} with PRF control and alternation constraint: $S_a = 1$.

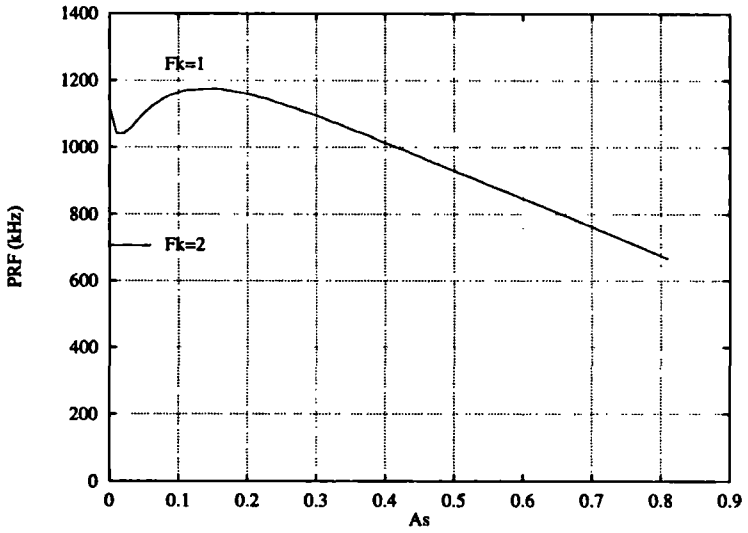


Figure 6.18: Average PRF against input level (1 kHz sinewave) for bit-flipping modulator {64, 4, 2.0} with PRF control and alternation constraint: $S_a = 2$

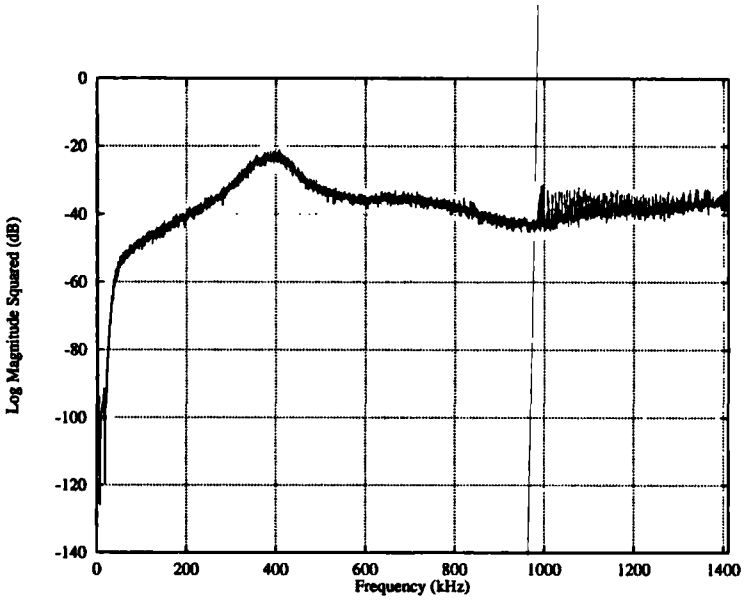


Figure 6.19: Wideband spectrum of output of bit-flipping algorithm for modulator {64, 4, 1.0} with PRF constraint and alternation constraint $S_a = 1, F_k = 5$

Definition 6.2 *N_{min} is defined as the minimum number of consecutive equi-signed samples observed at the input to the BFO.*

For example, the periodic sequence $b(k) = \overline{-1,-1,+1}$ has the value $N_{min} = 1$ and the periodic sequence $q(k) = \overline{-1,-1,-1,+1,+1}$ has the value $N_{min} = 2$. The definition of N_{min} also applies to aperiodic sequences. Some example sequences and their associated values of N_{min} are shown in figure 6.20. Also shown are the bit sequences resulting from the application of the bit-flipping algorithm with values of the alternation constraint $S_a = 1, 2$.

In these example patterns, the observation can be made that a *delay of S_a samples* occurs through the bit-flipping operator for all cases where $S_a < N_{min}$.

The value of N_{min} depends on the bit-flipping activity of the modulator. This is because errors introduced by the BFO are fed back and modify the future bit patterns of $b(k)$. The modifications can be considered to be bit-flips on the original $b(k)$. These are termed *intrinsic* bit-flips, as opposed to *algorithmic* bit-flips, which are directly produced by the bit-flipping operator.

To illustrate this, the bit sequence at the output of a first order $\Sigma\Delta$ modulator (figure 6.21) with zero input has been evaluated, firstly with the BFO inactive, then with the BFO active using the alternation algorithm with $S_a = 1$. The sequences are shown in table 6.6. In these sequences the quantizer is defined such that $b(k) = 1$ for $u(k) = 0$. With the BFO active, algorithmic bit-flipping occurs on every third

	N_{min}=2	b(k)	1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
(a)	S_a=1	y(k)	1	1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1
	S_a=2	y(k)	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	1	1	1	1
	N_{min}=3	b(k)	1	1	-1	-1	-1	1	1	1	1	-1	-1	-1	1	1	1	-1	-1
(b)	S_a=1	y(k)	1	1	1	-1	-1	-1	1	1	1	1	-1	-1	-1	1	1	1	-1
	S_a=2	y(k)	1	1	1	1	-1	-1	-1	1	1	1	1	-1	-1	-1	1	1	1

Figure 6.20: Example input bit sequence to BF operator with $N_{min} = 2, 3$ and the resulting output sequence for $S_a = 1, 2$

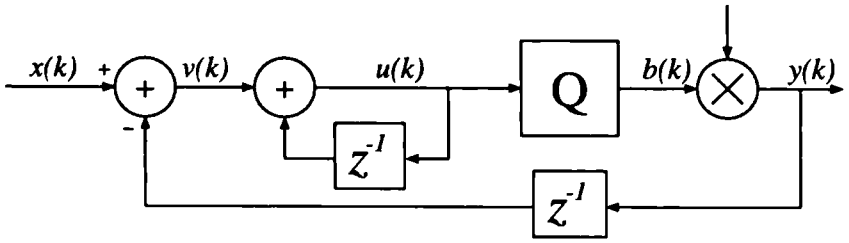


Figure 6.21: First order modulator with bit-flipping

sample. Samples of $b(k)$ become grouped. The difference between the values of $b(k)$ with and without bit-flipping are due to the intrinsic bit-flips. It can be seen that an intrinsic $-1 \rightarrow 1$ flip follows every algorithmic $1 \rightarrow -1$ flip and vice-versa. In section 6.5 exhaustive simulations are presented which show that this is a general result for a $\Sigma\Delta$ modulator which produces the alternating periodic pattern $\overline{1, -1}$.

As a result of the intrinsic bit-flipping, the value of N_{min} increases from 1 to 3. For $S_a = 1$, since $S_a < N_{min}$, the BFO can be modelled as a delay of one sample. In general, simulations have shown that increasing S_a causes N_{min} to increase, therefore the condition $S_a < N_{min}$ is satisfied and the BFO can be modelled as a delay of S_a samples. The introduction of the delay is consistent with the grouping of bits at the quantizer output in the sense that the delay prevents the feedback

No Bit-Flipping			Bit-Flipping				Type	
$v(k)$	$u(k)$	$y(k)$	$u(k)$	$v(k)$	$b(k)$	$y(k)$	A	I
0	0	1	0	0	1	1		
-1	-1	-1	-1	-1	-1	1	x	
1	0	1	-1	-2	-1	-1		x
-1	-1	-1	1	-1	-1	-1		
1	0	1	1	0	1	-1	x	
-1	-1	-1	1	1	1	1		x
1	0	1	-1	0	1	1		
-1	-1	-1	-1	-1	-1	1	x	
1	0	1	-1	-2	-1	-1		x
-1	-1	-1	1	-1	-1	-1		
1	0	1	1	0	1	-1	x	
-1	-1	-1	1	1	1	1		x
1	0	1	-1	0	1	1		
...								

Table 6.6: Variable states of 1st order modulator with zero input: (a) No bit-flipping, (b) Bit-flipping, $S_a = 1$. Also shown is the type of bit-flipping: A=algorithmic, I-intrinsic.

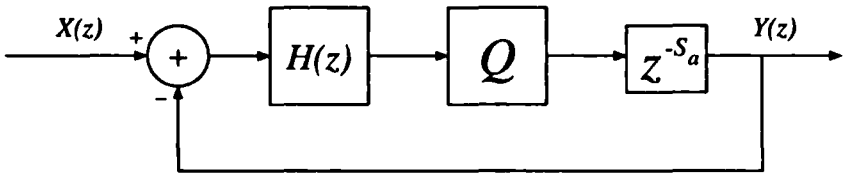


Figure 6.22: Modelling alternation constraint as a modulator with additional unit delays

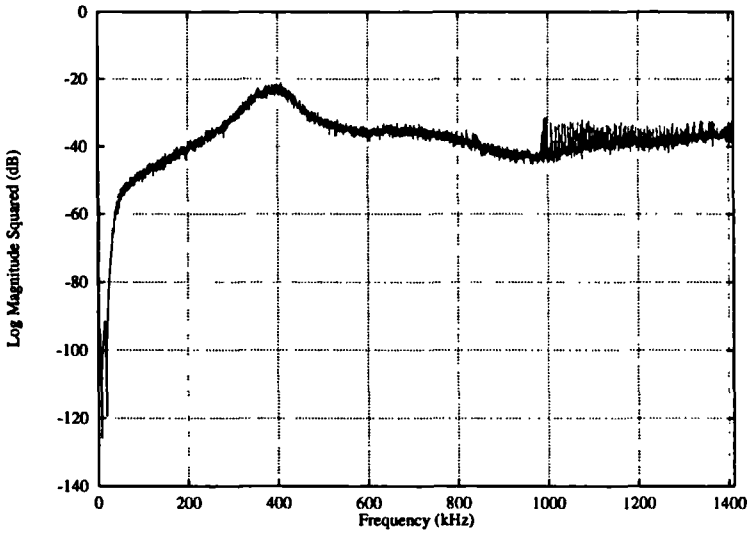


Figure 6.23: Wideband spectrum of modulator {64, 4, 1.0} with additional unit delay for parameter $A_s = 0.1$

responding immediately to the quantization error and so prevents the occurrence of high frequency oscillating limit cycles.

To demonstrate the validity of this approach, a simulation has been performed of a modulator with an additional delay with parameters $L = 64, P_n = 1\ dB$ and a $1\ kHz$ input of amplitude $A_s = 0.1$. The wideband spectral response of this system is shown in figure 6.23. The response is identical to figure 6.19,which is the result from a simulation of a modulator employing bit flipping with $S_a = 1$ and $P_n = 1\ dB$. For this example, the modelling is valid. The reader is referred to appendix E.3 for a confirmation of the model for a wide range of simulation parameters.

6.4.1 Mapping Bit-Flipping onto the NTF

Since the bit-flipping algorithm with alternation factor S_a can be modelled as a simple delay of S_a samples, it is possible to map the bit-flipping onto an equivalent

operation on the modulator NTF. For a loop-filter $H(z)$, the NTF can be expressed as

$$NTF_d(z) = \frac{1}{1 + H(z)z^{-S_a}} \tag{6.40}$$

where $H(z)$ contains a single delay by implication i.e. $H(z) = z^{-1}H'(z)$ and $H'(z)$ is a delay-free transfer function

We now consider how the NTF poles and zeros are affected by this delay. For a modulator NTF with no additional delay separated into poles and zeros $NTF(z) = A(z)/B(z)$:

$$H(z) = \frac{B(z) - A(z)}{A(z)} \tag{6.41}$$

Hence from equation 6.40

$$NTF_d(z) = \frac{A(z)}{A(z)(1 - z^{-S_a}) + B(z)z^{-S_a}} \tag{6.42}$$

The delay affects the poles but not the zeros of $NTF(z)$. The effect is considered initially for a first order modulator with $S_a = 1$:

$$NTF(z) = 1 - z^{-1} \tag{6.43}$$

$$NTF_d(z) = \frac{1 - z^{-1}}{1 - z^{-1} + z^{-2}} \tag{6.44}$$

Pole locations are introduced at locations $z_1 = 0.5 - j\frac{\sqrt{3}}{2}$ and $z_2 = z_1^* = 0.5 + j\frac{\sqrt{3}}{2}$, equivalent to a complex conjugate pole on the unit circle at frequency $\theta = \pi/3$. This causes a resonance in the NTF at a frequency $Lf_s/6$. In figure 6.24 the wideband spectral response of a first order modulator with an additional unit delay is plotted. The parameters are $L = 64$, $A_s = 0.2$. A peak in noise spectrum occurs at a frequency of 468.6 kHz which is close to the resonant frequency of 470.4 kHz predicted by the model. The average PRF of the bitstream occurs close by at 483.3 kHz. The proximity of the resonant frequency and the average PRF of the bitstream has also been suggested in [Agr83]. A possible explanation is that any periodic components in the bitstream will occur as spectral peaks in the frequency domain. The noise-like structure of the spectrum indicates that the periodic bitstream frequency is modified randomly i.e. the short-term PRF randomly deviates from the average PRF.

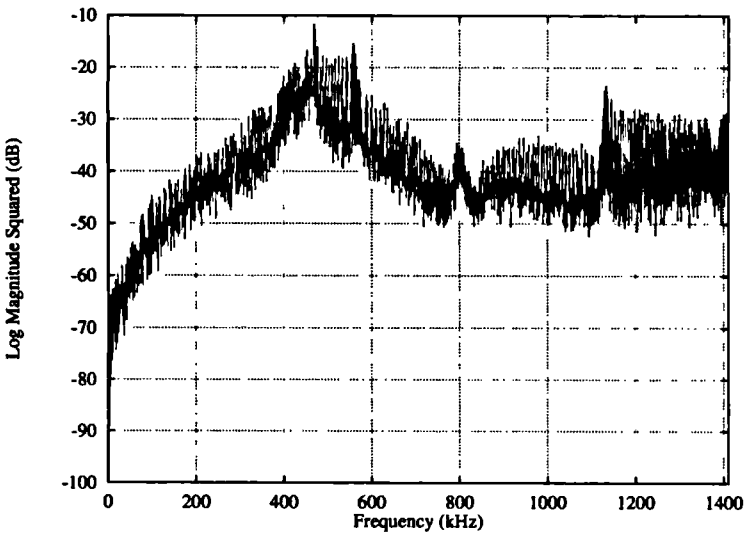


Figure 6.24: Wideband spectrum of first order modulator with additional unit delay

6.4.2 Extension to Higher Order and Quasi-linear Model

In this section the analysis is extended to higher order systems and it is shown how the quasi-linear quantizer model can be used to achieve fairly accurate PRF predictions and account for a change in PRF which occurs with input level. For a general order modulator with delay S_a and quasi-linear noise gain K_n the NTF is given by:

$$NTF_{dk}(z) = \frac{1}{1 + H(z)K_n z^{-S_a}} \tag{6.45}$$

Again expressing the transfer functions in terms of the unmodified transfer function poles and zeros: $NTF(z) = A(z)/B(z)$:

$$NTF_{dk}(z) = \frac{A(z)}{A(z)(1 - K_n z^{-S_a}) + B(z)K_n z^{-S_a}} \tag{6.46}$$

The delay and gain terms affect only the poles of the NTF. To demonstrate this, the z-plane pole locations for the modulator {64, 4, 1.0} for delays $S_a = 0, 1, 2$ with a gain of $K_n = 1$ is plotted in figures 6.25 and 6.26. The additional delays cause the effective order of the modulator to increase by S_a, resulting in additional poles on the z-plane. The location of the existing poles also changes. The overall effect is to modify the frequency response of the NTF. In figure 6.27, the frequency response of the NTF with $S_a = 0, 1, 2$ and $K_n = 1$ is plotted. The delay causes resonance in the NTF. The resonance increases the power gain of the filter and this causes

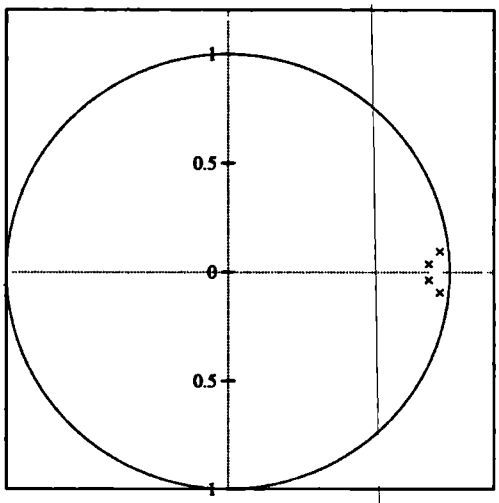


Figure 6.25: z-plane pole constellation for the modulator {64, 4, 1.0} with $K_n = 1$, $S_a = 0$

the stability of the modulator to deteriorate. The increase in power gain is shown in figure 6.28(a) in which the power gain with delay $S_a = 1$ and $S_a = 2$ is plotted against the power gain with no delay.

Figure 6.27 also shows that the baseband response is insensitive to modification to the pole locations. This leads to the surprising result that if K_n were to remain constant, the added loop delays would *not* affect the noise performance of the modulator.

In practical operation, the quasi-linear gain is not unity, but varies with modulator input level. The values of K_n for the modulators with $S_a = 0, 1, 2$ and $P_n = 1\ dB$ have been obtained by simulating a real modulator over its stable input range using the quasi-linear time-average method of section 3.3.2 with 100,000 simulation samples. The values are plotted in figure 6.28(b). Notice that K_n reduces with S_a. This is because the delay prevents the loop from quickly correcting quantization errors and so the error at the modulator input node increases, causing the variance of the quantizer input signal $u(k)$ to increase and K_n to fall. The increase in the variance of $u(k)$ is also consistent with the increase in power gain of the filter, which increases the level of noise circulating in the loop. The fall in K_n causes a reduction in the modulator stability (observed in figure 6.28(b) by a reduced maximum input level) and an increase in baseband noise.

The variation of the pole locations with changing K_n can be plotted using the

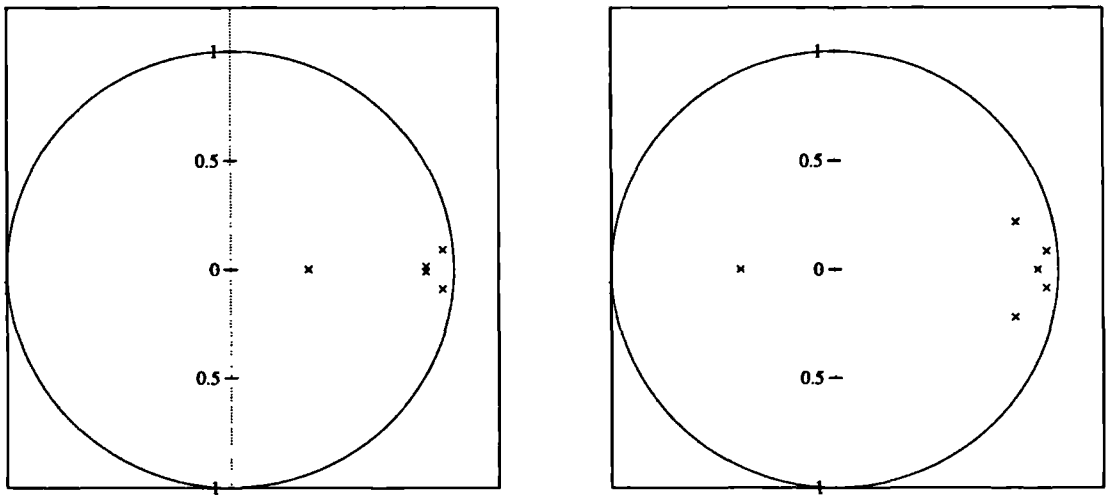


Figure 6.26: z-plane pole constellation for the modulator {64, 4, 1.0} with $K_n = 1$, (a) $S_a = 1$, (b) $S_a = 2$

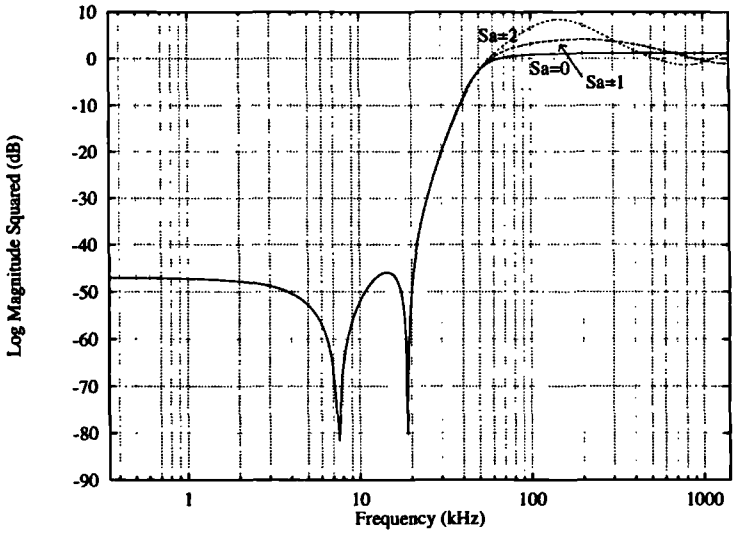
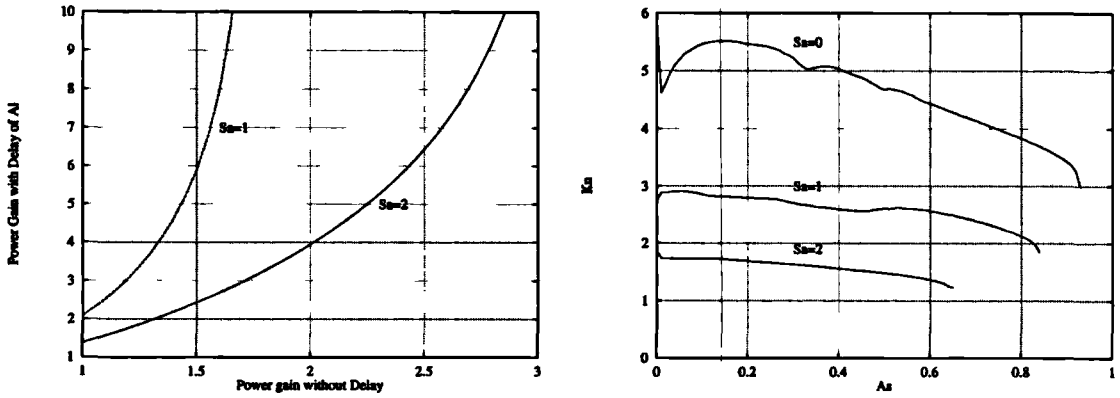


Figure 6.27: NTF frequency response for delaying $\Sigma\Delta$ modulator {64, 4, 1.0} with $P_n = 1$, $K_n = 1$, $S_a = 0, 1, 2$.



(a) (b)

Figure 6.28: (a) NTF power gain with additional delay S_a against power gain with no additional delay, for $L = 64$, $N = 4$, $K_n = 1$, $S_a = 1, 2$. (b) Variation in K_n with sinusoidal input level for the modulator {64, 4, 1.0} with $S_a = 0, 1, 2$

root-locus technique. This technique has been used used previously [Sti88b] for estimating stability using quasi-linear analysis. In figure 6.29 the root locus of the NTF is plotted for K_n in the range $1 \rightarrow 4$ and $S_a = 1$ and 2. The range of K_n observed in a modulator simulation with a 1 kHz sinewave input, over its stable amplitude range are as follows:

$$S_a = 1: \quad K_n = 1.19 \rightarrow 1.34$$
$$S_a = 2: \quad K_n = 1.10 \rightarrow 1.24.$$

The pole locations corresponding to this range are plotted in the root locus using heavy type. It can be seen that some of the complex conjugate poles are considerably more sensitive to variations in K_n than others. The sensitive poles on the right hand side of the z-plane correspond in frequency to a peak in the noise transfer function and will be referred to as the *dominant* poles. For zero input, these poles are located close to the unit circle. As the input amplitude decreases, K_n decreases and the dominant poles move inwards, away from the unit circle and the pole frequency decreases. As an example of the relationship between the frequency response of the NTF using the quasi-linear model and the noise spectrum, the value of K_n has been determined by simulating the modulator with a 1 kHz sinewave input, for $A_s = 0.1$ and using this value in the NTF frequency response. There is a strong correspondence between the peak in the NTF (figure 6.30) and the peak in the noise spectrum of the simulated modulator (figures 6.31). The differences are

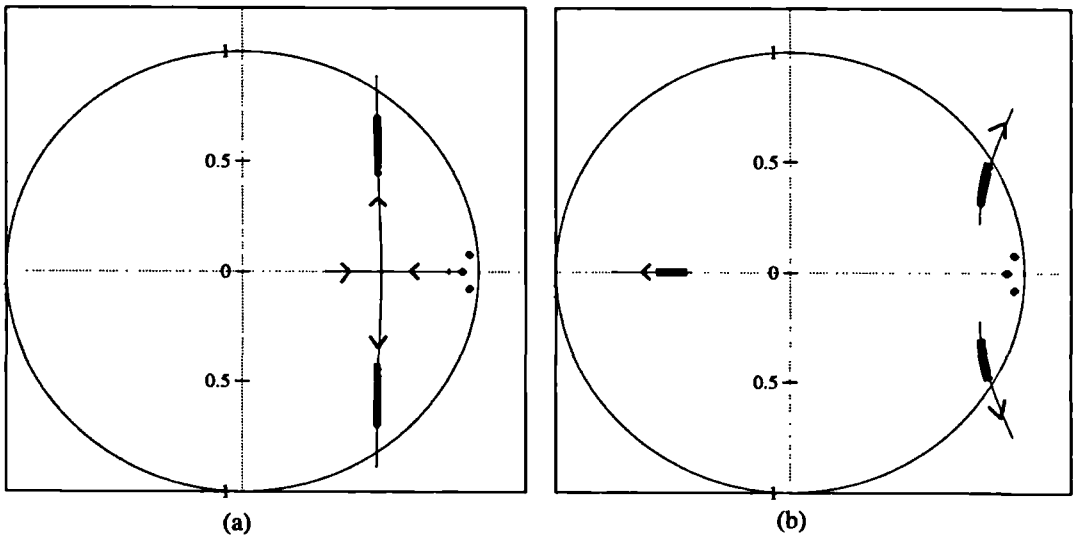


Figure 6.29: Root Locus of the NTF of modulator {64, 4, 1.0} with (a) delay of $S_a = 1$, (b) delay of $S_a = 2$

most probably attributable to a non-white quantization error.

The resonance in the noise spectrum strongly influences the average PRF of the modulator. In figure 6.32 the dominant pole frequencies are compared to the modulator PRF for $S_a = 1$ and $S_a = 2$. There is a close correspondence, especially for $S_a = 2$, where the resonance is stronger and dominates the PRF.

6.4.3 Summary

In this section it has been shown that the alternation constraint can be modelled accurately as a $\Sigma\Delta$ modulator with additional loop delays. The delays modify the NTF, causing an out-of-band resonance. The quasi-linear model has been used to model the variation in resonant frequency with input level. It has been shown that this resonant frequency corresponds fairly closely to the average PRF of the bitstream, The delays also cause the noise and stability of the modulator to suffer, due to a reduction in quasi-linear gain.

6.4.4 Using Delaying Sigma-Delta modulators for Power D-A Conversion

The results presented for modulators with additional delays in the modulator loop show that a significant reduction in PRF is possible, in comparison to a standard

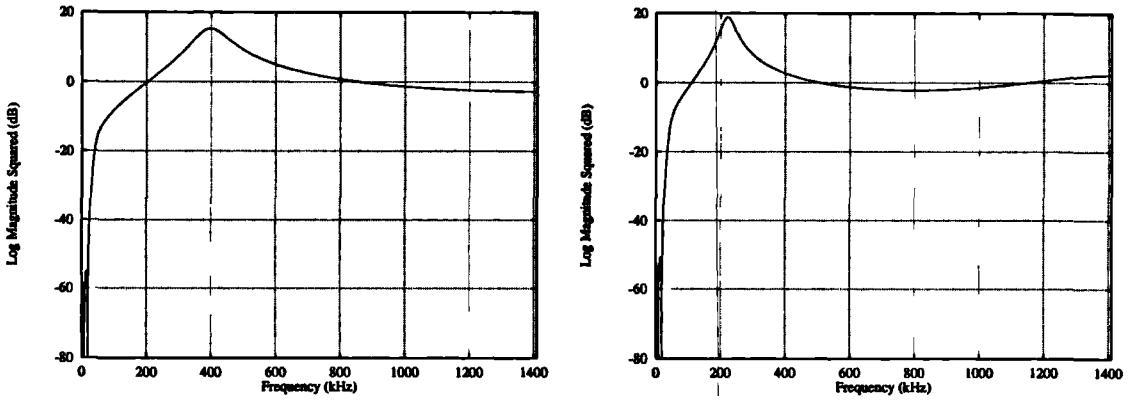


Figure 6.30: NTF frequency response of modulator {64, 4, 1.0} with (a) $S_a = 1$, $K_n = 2.85$ (b) $S_a = 2$, $K_n = 1.73$

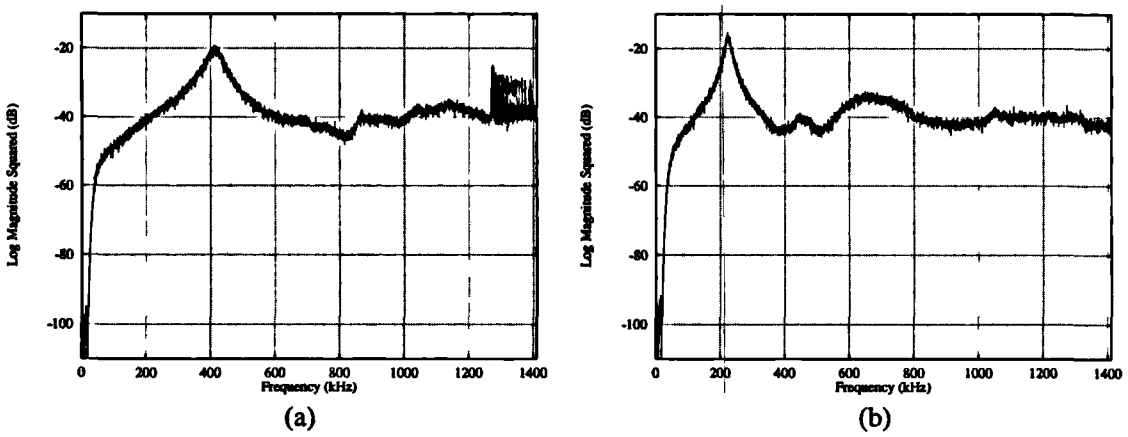


Figure 6.31: Spectral response of modulator {64, 4, 1.0} with (a) $S_a = 1$, (b) $S_a = 2$

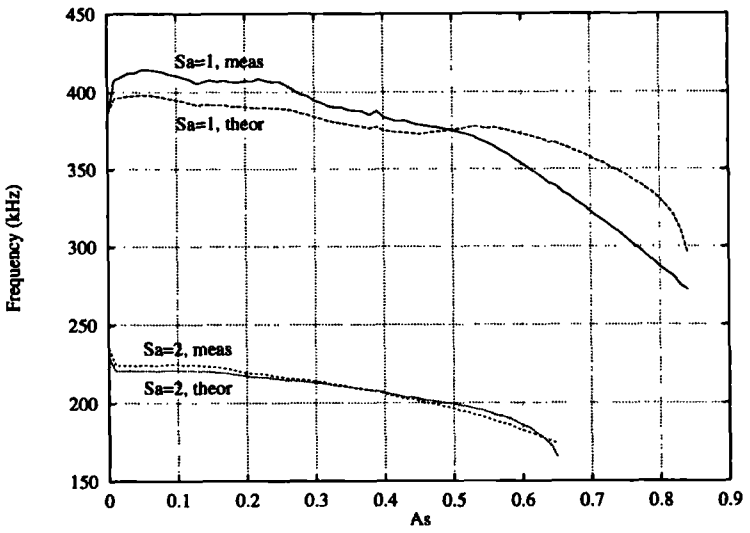


Figure 6.32: PRF of ΣΔ modulator {64, 4, 1.0} with additional unit delay against input amplitude A_s, measured and theoretical

modulator. This leads to the possibility of using delaying modulators in power D-A converters, with a reduction in complexity over the bit-flipping algorithm. In section 6.6.5 results will be presented to compare the performance of delaying ΣΔ modulators with different bit-flipping algorithms. It will be shown that, although the dynamic range of the delaying modulator is almost identical to the bit-flipping modulator with alternation constraint and PRF control, the delaying system suffers inferior performance when used with a non-ideal power switching stage. Furthermore, the large peak in the out of band noise at the average PRF is undesirable because nonlinearities following the sample and hold may cause intermodulation of the noise back into the baseband [Nau91]. In conclusion, although the idea of using a delaying modulator is attractive due to its low relative complexity, for practical modulators the technique is undesirable.

6.5 Look-Ahead

It has been shown that algorithmic and intrinsic bit-flipping can introduce a delay through the bit-flipping unit. The delay modifies the NTF, causing the stability and noise performance to suffer. In this section a bit-flipping algorithm is investigated in which delay caused by algorithmic bit-flipping is prevented. The method is to precisely calculate possible future bit patterns to allow a more informed decision

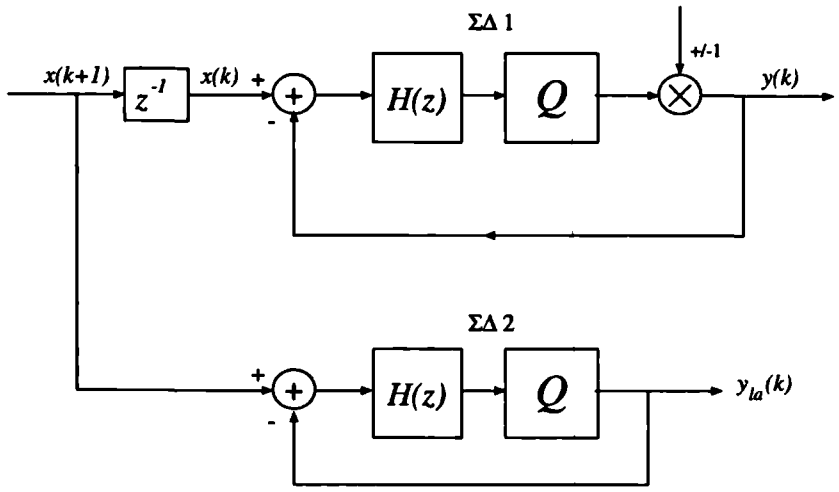


Figure 6.33: Block diagram of conceptual look-ahead system

to be made by the bit-flipping unit. It will be shown that the stability and noise properties of the modulator can be considerably improved in this way. The technique is termed *look-ahead*. The specific examples chosen use the look-ahead algorithm in conjunction with the alternation constraint and PRF control, though it will be shown by simulation in section 6.6.6 that it can be applied also to the quantizer input bound. In developing the algorithm, examples are used in which it is assumed only algorithmic bit-flipping occurs. The operation will then be examined for both intrinsic and algorithmic bit-flipping.

The principle of look-ahead has been used in [Cra93] to prevent delays in a PWM feedback correction algorithm. The basic idea is to delay the input to the modulator and use the resulting 'advance' input to feed an identical modulator (ΣΔ2 in figure 6.33). The look-ahead output of ΣΔ2 ($= y_{la}(k)$) is therefore a prediction of the next output of ΣΔ1 ($y(k+1)$). The idea of look-ahead is to use the advance output to influence the decision made by the BFO.

6.5.1 Violation of Causality

There are two fundamental problems with this simple scheme. Firstly, it is clear from figure 6.33 that any bit-flipping of ΣΔ1 will cause the predictions made by ΣΔ2 to be incorrect in subsequent samples. Secondly, a violation of causality is provoked by attempting to use a prediction of $y(k+1)$ to control the BFO which produces $y(k)$.

The non-causal loop can be broken by calculating $y_{la}(k)$ for the case where no bit-flipping occurs in sample k. It will be shown in section 6.5.2 that this information is sufficient to improve the performance of the algorithm. If bit-flipping does occur, $y_{la}(k)$ is recalculated so that the predictions made by $\Sigma\Delta 2$ will be correct in subsequent samples. Full details of an efficient implementation of this algorithm is presented in appendix E.4.

6.5.2 The Look-Ahead Algorithms

To understand how look-ahead can be used to improve the operation of the bit-flipping algorithm, consider again the examples of figure 6.20. Where $S_a < N_{min}$, the bit-flipping is unsuccessful at *directly* reducing the transitions in the bitstream because the alternation factor prevents the flipping of enough consecutive bits to reduce the transition rate (for the moment the possibility of intrinsic bit-flipping is ignored). This means that there is unnecessary bit-flipping which does not directly contribute to reducing the transition rate. The bit-flipping causes a delay which degrades the modulator stability and noise performance. With knowledge of the next quantizer output, however, the bit-flipping can be restricted to the cases where there will be a definite reduction in transitions.

One-level of Look-Ahead

The simplest look-ahead algorithm uses one level of look-ahead (1LA). In this algorithm the current output $y(k)$ is used in conjunction with the previous output $y(k-1)$ (a delayed version of $y(k)$) and look-ahead output $y_{la}(k)$ to identify the following high transition rate patterns:

$$\{y(k-1), y(k), y_{la}(k)\}=\{1, -1, 1\}$$
$$\{y(k-1), y(k), y_{la}(k)\}=\{-1, 1, -1\}.$$

An attempt is then made to reduce the transitions by inverting $y(k)$. It is of course possible that an intrinsic bit-flip will follow the algorithmic bit-flip, therefore with reference to table 6.7 there are two possible bit-triplets which may occur. Notice that in both cases number of transitions T is reduced. The value of T is calculated for transitions within the pulse group (i.e. the assumption is made that there are no transitions immediately before or after the group). By bit-flipping *only* on the

Before BF		After BF	
$y(k-1)$, $y(k)$, $y_{1a}(k)$	T	$y(k-1)$, $y(k)$, $y(k+1)$	T
-1, 1, -1	2	-1, -1, -1	0
		-1, -1, 1	1

Table 6.7: Possible patterns before and after bit-flipping for $LA = 1$

occurrence of the two high transition rate patterns {1, -1, 1} and {-1, 1, -1}, a guaranteed reduction in transitions occurs for every bit-flip.

Two Levels of Look-ahead

In 'reducing' (i.e. reducing the number of transitions) only the {-1, 1, -1} or {1, -1, 1} triplets , there is a limit to the reduction in PRF possible.

In the two-level look-ahead (2LA) algorithm an additional output $y_{2la}(k)$ is generated by the modulator, representing a prediction of $y(k+2)$ for the case where no bit-flipping occurs in sample $y(k)$ or $y(k+1)$. This makes it possible to predict and eliminate the occurrence of the following patterns.

1. $\{y(k-1), y(k), y_{1a}(k)\} = \{1, -1, 1\}$
2. $\{y(k-1), y(k), y_{1a}(k)\} = \{-1, 1, -1\}$
3. $\{y(k-1), y(k), y_{1a}(k), y_{2la}(k)\} = \{1, -1, -1, 1\}$
4. $\{y(k-1), y(k), y_{1a}(k), y_{2la}(k)\} = \{-1, 1, 1, -1\}$

This enables the PRF to be reduced further than with the 1LA algorithm.

In the case of bit-patterns 3 and 4, assuming no intrinsic bit-flipping occurs, two adjacent bit-flips are required i.e. both $y(k)$ and $y(k+1)$ need to be inverted. This is accommodated naturally by the algorithm, since inverting $y(k)$ in pattern 3 or 4 will produce pattern 1 or 2 on the next sample.

We now consider the possible bit-patterns which can occur after bit-flipping, taking into account intrinsic bit-flipping. There are six possible patterns of $y(k-1)$, $y(k)$, $y_{1a}(k)$, $y_{2la}(k)$ for which bit-flipping is allowed, and eight possible patterns at the output if bit flipping occurs. Half of these are shown in table 6.8 and the remainder are their inverses. Each combination has an associated reduction in transitions, however notice that there are two patterns (shown in bold type) where there is no reduction in transitions and so the occurrence of these patterns should be avoided.

Input		Output	
$y(k-1), y(k), y_{1a}(k), y_{2la}(k)$	T	$y(k-1), y(k), y(k+1), y(k+2)$	T
-1, 1, -1, 1	3	-1, -1, -1, -1	0
		-1, -1, -1, 1	1
		-1, -1, 1, -1	2
		-1, -1, 1, 1	1
-1, 1, -1, -1	2	-1, -1, -1, -1	0
		-1, -1, -1, 1	1
		-1, -1, 1, -1	2
		-1, -1, 1, 1	1
-1, 1, 1, -1	2	-1, -1, -1, -1	0
		-1, -1, -1, 1	1
		-1, -1, 1, -1	2
		-1, -1, 1, 1	1

Table 6.8: Possible patterns at the input and output of the bit-flipping unit for $LA = 2$. Only detected input patterns are shown.

6.5.3 Evaluation of Intrinsic Bit-flipping

A method of modifying the look-ahead algorithm to avoid bit-flipping where there is no significant reduction in transitions would be to evaluate intrinsic bit-flipping by calculating the future bit-patterns resulting from bit-flipping $y(k)$. In this scheme, addition look-ahead outputs would be generated for the case where bit-flipping occurs and these would be used to steer the bit-flipping algorithm. To establish whether such a modification could significantly improve the performance of the look-ahead algorithm, we need to investigate the probabilities of the different bit-patterns occurring.

This has been done using a number of test simulations of the modulator using different parameters and input signals. The modulator is simulated until bit-flipping occurs. When bit-flipping of $y(k)$ occurs, the next two output samples $y(k+1)$ and $y(k+2)$ are compared to the values predicted with no bit-flipping: $y_{1a}(k)$ and $y_{2la}(k)$. There is a counter for each combination of $y(k-1)$, $y(k)$, $y(k+1)$, $y(k+2)$, $y_{1a}(k)$ and $y_{2la}(k)$. The counter is incremented when each combination occurs.

The test simulations have been performed for modulators using the bit-flipping algorithm with alternation constraint and two-levels of look ahead. The simulation

Sequence	$y(k-1)$	$y(k)$	$y_{1a}(k)$	$y_{21a}(k)$	$y(k+1)$	$y(k+2)$
1	-1	1	1	-1	1	1
2	-1	1	-1	1	1	1
3	-1	1	-1	-1	1	1
4	-1	1	-1	1	1	-1

Table 6.9: Sequences which occur in intrinsic bit-flipping simulations

parameters are as follows:

- Oversampling ratio $L = 64$
- Orders $N = 3$ to $N = 8$
- Power gains $P_n = 1 \rightarrow 3$ in steps of 0.5
- Alternation constraint $S_a = 1, 2$
- 1 kHz sinewave input of input amplitude A_s in the range $0 \rightarrow 0.5$ in steps of 0.1.

For simulations of 1 million samples, it has been found that there are only eight combinations of outputs which occur, and four of these are given in table 6.9. The other combinations are the inverses of these patterns (i.e. every 1 is replaced by a -1 and vice-versa). Note that a necessary condition for algorithmic bit-flipping to occur is $y(k-1) \neq y(k)$ (refer to the basic algorithm flow chart of figure 6.8) . The values of $y(k-1)$ are also shown in table 6.9.

The number of occurrences of each pattern are shown in tables 6.10 and 6.11 for the values of alternation constraint $S_a = 1$ and $S_a = 2$, respectively. Although these results have been generated for the 2LA algorithm, they are also applicable to the 1LA algorithm by considering only the 1LA outputs. These results are included only as evidence of the simulations, since it is the combinations which have zero occurrences (not shown) which are important.

Looking initially at the 1LA algorithm, for the pattern which is detected by the algorithm: $y(k-1)$, $y(k)$, $y_{1a}(k)$={1, -1, 1}, the possible sequences are 2,3, 4. All of these sequences have an intrinsic bit-flip occurring on sample $k+1$, therefore referring to table 6.7, only the lower sequence can occur, which causes a reduction of only one transition. The measured probability of occurrence of the upper pattern is

		Sequence			
P_n	A_s	1	2	3	4
1.00	0.00	3036.9	21361.7	1849.6	0.0
1.00	0.10	3665.4	22183.9	2602.9	56.6
1.00	0.20	4901.4	19363.0	3292.0	1275.0
1.00	0.30	4265.0	16542.4	3189.3	2456.0
1.00	0.40	5025.7	14071.1	3197.6	4332.4
1.00	0.50	4611.0	12395.1	2401.7	6175.4
1.50	0.00	3342.9	21680.1	1411.9	0.0
1.50	0.10	3709.4	19526.1	2735.7	2.9
1.50	0.20	4894.9	17198.4	3354.9	347.4
1.50	0.30	5456.6	15638.0	3629.7	1084.4
1.50	0.40	5482.0	13075.6	3722.9	2156.1
1.50	0.50	4864.1	10728.7	3349.3	2784.4
2.00	0.00	7602.0	13669.7	1884.6	0.0
2.00	0.10	6354.7	15741.7	2502.7	5.6
2.00	0.20	6194.1	14883.7	2759.1	65.6
2.00	0.30	5844.4	13639.4	3232.0	324.7
2.00	0.40	5582.3	11950.1	3449.0	754.0
2.00	0.50	5095.6	9915.9	3234.6	1122.4
2.50	0.00	7286.4	12071.6	1138.6	1.6
2.50	0.10	6740.1	12056.0	1689.0	6.6
2.50	0.20	6410.6	11726.9	1873.7	32.0
2.50	0.30	4632.9	10786.6	2090.4	92.4
2.50	0.40	4362.7	9837.3	2099.3	161.7
2.50	0.50	5932.0	8027.0	1966.50	186.0
3.00	0.00	4161.3	13873.7	1.7	0.3
3.00	0.10	4547.0	9032.0	670.50	14.50
3.00	0.20	3210.5	9067.0	811.0	12.0
3.00	0.30	6377.5	8629.50	1072.0	45.50
3.00	0.40	6947.0	9044.0	1463.0	0.0
3.00	0.50	6425.0	7563.0	1719.0	14.0

Table 6.10: Occurrences of bit combinations for $S_a = 1$

		Sequence			
P_n	A_s	1	2	3	4
1.00	0.0	11204.7	17887.7	719.0	0.0
1.00	0.1	11408.7	17177.9	1458.6	30.6
1.00	0.2	8558.7	16805.3	2180.4	713.7
1.00	0.3	8003.9	14637.6	2635.4	1752.3
1.00	0.4	7421.7	12733.1	2809.1	3873.4
1.00	0.5	5619.7	11376.3	2079.0	5885.1
1.50	0.0	6257.0	18661.6	932.6	0.0
1.50	0.1	8188.9	16081.6	1586.4	14.4
1.50	0.2	8381.7	15520.6	2383.6	168.6
1.50	0.3	7417.7	14332.9	3116.6	868.0
1.50	0.4	5076.4	12240.3	3313.3	2036.3
1.50	0.5	5890.6	10031.1	3053.1	2711.1
2.00	0.0	9111.0	13189.6	1483.3	13.4
2.00	0.1	8785.7	13429.1	1651.1	36.1
2.00	0.2	7733.9	13523.6	2182.0	84.3
2.00	0.3	6816.3	12796.7	2824.0	335.9
2.00	0.4	6284.0	11318.1	3164.6	753.0
2.00	0.5	4224.7	9410.4	3018.4	1098.7
2.50	0.0	7977.1	11061.9	1042.1	28.4
2.50	0.1	7678.6	10788.9	1390.9	41.7
2.50	0.2	7182.4	10879.2	1452.4	54.2
2.50	0.3	6755.0	10258.5	1662.0	75. 3
2.50	0.4	6619.0	9540.5	1888.0	68.5
2.50	0.5	6019.0	8035.5	1887.5	165.5
3.00	0.0	3803.0	11722.5	7.5	0.0
3.00	0.1	4901.5	8694.5	639.0	19.0
3.00	0.2	7236.5	8774.0	757.5	40.0
3.00	0.3	6743.5	8481.0	976.0	48.5
3.00	0.4	7214.0	8762.0	1422.0	0.0
3.00	0.5	6582.0	7404.0	1662.0	25.0

Table 6.11: Occurrences of bit-combinations for $S_a = 2$

zero. This means that, for the 1LA algorithm, every $\{1 \rightarrow -1\}$ flip is followed by an intrinsic $\{-1 \rightarrow 1\}$ flip and vice-versa. As a consequence the maximum reduction in transitions never occurs and therefore there is *no* advantage in modifying this algorithm.

For the 2LA algorithm, all the patterns which occur correspond to a reduction in transitions (refer to table 6.8). The patterns in table 6.8 in which there is no reduction in transitions do not occur. Also the patterns for which there is a maximal reduction in transitions do not occur.

In conclusion, there would be no advantage in modifying the operation of the look-ahead algorithms since a reduction in transitions will occur for every bit-flip and of the bit-patterns which occur, there are none which should be avoided.

6.5.4 Examples

We now briefly consider the effectiveness of the look-ahead algorithm in conjunction with the PRF control for two NTF power gains. Further results for the look-ahead in conjunction with the alternation constraint and quantizer input bound will be given in section 6.6.6. Simulations have been performed to evaluate the PRF versus input amplitude for a 1 kHz sinewave input for the modulator {64, 4, 2.0}. In figures 6.34 and 6.35 the results are plotted for one and two levels of look-ahead with values of F_k simulated in the range $F_k = 1 \rightarrow 5$. It can be seen that the look-ahead has a similar effect to the alternation constraint - a lower bound is introduced onto the PRF and the stability of the modulator is improved. The lower bound is at a slightly higher PRF than with the alternation constraint (figure 6.17), however this restriction is relaxed for $LA = 2$. Compared to the alternation constraint, the stability of the modulator is improved using look-ahead.

6.6 Investigations and Results

In this section the following algorithms are compared in detail.

- System A: 'Standard' $\Sigma\Delta$ system.
- System B: Bit-flipping algorithm with PRF control only.
- System C: Bit-flipping algorithm with PRF control and quantizer input bound.

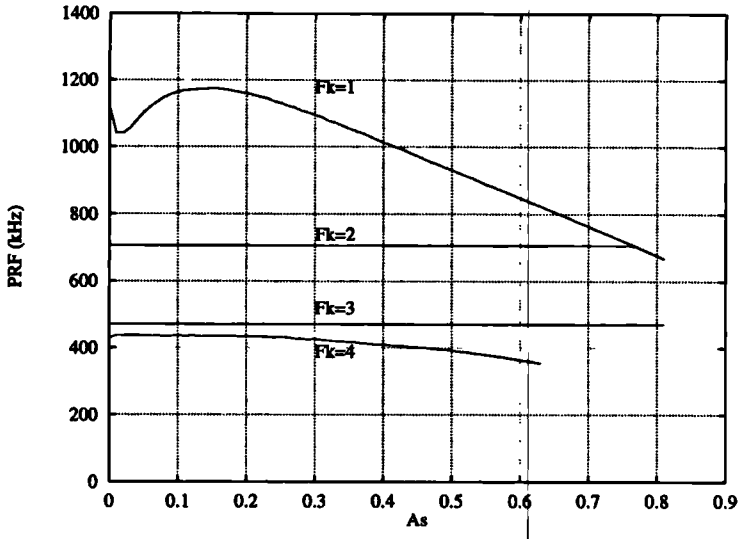


Figure 6.34: Average PRF against input level (1 kHz sinewave) for bit-flipping modulator {64, 4, 2.0} with PRF control and one level of look-ahead.

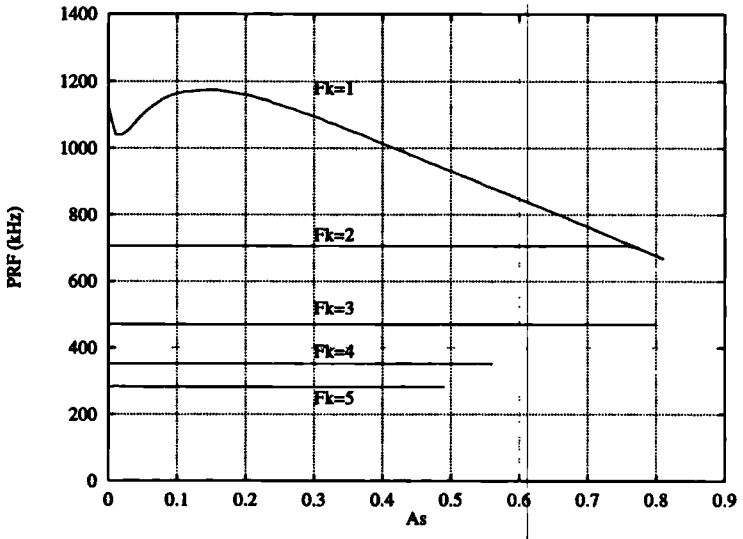


Figure 6.35: Average PRF against input level (1 kHz sinewave) for bit-flipping modulator {64, 4, 2.0} with PRF control and two levels of look-ahead.

- System D: Bit-flipping algorithm with PRF control and alternation constraint.
- System E: $\Sigma\Delta$ with additional loop delays.
- System F: Bit-flipping algorithm with look-ahead and PRF control.
- System G: Bit-flipping algorithm with look-ahead, quantizer input bound and PRF control.
- System H: Bit-flipping algorithm with look-ahead, alternation constraint and PRF control.

Where the algorithms are combined, all of the conditions of the individual algorithms must be satisfied for the BFO to be triggered.

6.6.1 Evaluation of Relative performance

The examples used previously to evaluate the performance of the different bit-flipping algorithms (for example section 6.5.4) have shown that the stability and hence MSA of the modulators using bit-flipping depend upon the algorithm and parameters. It is difficult to usefully compare two algorithms with vastly different MSAs because the noise performance of a system optimised for high maximum input level is often considerably poorer than a system optimised for low maximum input level. Therefore all the modulators tested are first optimised to have the same MSA.

All the modulators described in the following are optimised to achieve a MSA which has been set at 0.3, which is within the range normally associated with optimal modulators. The NTF parameters to achieve this have been found according to the method of appendix A.4.

A simulation is performed at the maximum power gain P_m and input level to determine the baseband noise power P_b. PRF measurements are then made over the input range to determine the PRF constancy.

These tests are performed for different orders, oversampling ratios and bit-flipping parameters. For the $L = 32$ and $L = 64$ modulators, orders 4 to 8 are investigated. For the $L = 128$ modulator, where the dynamic range is considerably greater than the analogue signal processing would be able to achieve, orders 3 to 7 are investigated.

	L=32			L=64		
order	Max PRF(kHz)	P_m	$P_b(dB)$	Max PRF(kHz)	P_m	$P_b(dB)$
4	497.5	4.20	-92.2	993.9	4.25	-119.1
5	496.3	4.15	-100.2	1004.1	4.05	-132.0
6	504.6	4.10	-106.8	1003.5	4.00	-144.1
7	505.2	4.00	-111.1	1007.9	4.00	-154.5
8	502.8	4.05	-116.4	1013.4	3.95	-164.0

Table 6.12: Summary of Performance of Standard modulator (System A), $L = 32$ and $L = 64$

order	Max PRF(kHz)	P_m	$P_b(dB)$
3	1968.3	4.45	-123.9
4	1993.7	4.20	-145.9
5	2001.6	4.05	-165.0
6	2035.5	3.95	-182.3
7	2017.2	3.90	-197.6

Table 6.13: Summary of Performance of Standard modulator (System A), $L = 128$

Referring to section 6.1.1, the target performance for the systems to achieve is a PRF of 352.8 kHz and a dynamic range of 98 dB. For the maximum input of $A_s = 0.3$, the target baseband noise power is $98 + 10.5\ dB = 108.5\ dB$.

6.6.2 System A: Standard Sigma-Delta System

We begin by comparing the performance of standard $\Sigma\Delta$ modulators with different orders and oversampling ratios. The results are summarised in tables 6.12 and 6.13. The PRF is the maximum evaluated over the input range of the modulator. The results show that the target PRF of 352.8 kHz cannot be achieved for the oversampling ratios shown. To achieve the target PRF, a lower oversampling ratio is required, with a commensurate order increase to achieve the target baseband noise power. For power D-A converters, very high order filters are undesirable because of the increased complexity required in the high power passive reconstruction filter.

order	F_k	L=32 PRF(kHz)	P_m	$P_b(dB)$	L=64 PRF(kHz)	P_m	$P_b(dB)$
4	2	352.8	2.45	-70.2	705.6	2.35	-96.0
	3	235.2	1.65	-55.0	470.4	1.50	-79.7
	4	176.4	1.25	-45.1	470.4	1.50	-79.7
5	2	352.8	2.30	-56.4	705.6	2.25	-104.0
	3	235.2	1.60	-56.4	470.4	1.40	-84.3
	4	176.4	1.10	-44.5	352.8	1.05	-71.8
6	2	352.8	2.25	-75.1	705.6	2.10	-109.41
	3	235.2	1.60	-57.3	470.4	1.45	-89.5
	4	176.4	1.25	-45.6	352.8	1.05	-74.1
7	2	352.8	2.20	-76.0	705.6	2.10	-115.1
	3	235.2	1.60	-57.2	470.4	1.40	-91.5
	4	176.4	1.30	-44.8	352.8	1.00	-73.1
8	2	352.8	2.35	-79.3	352.8	1.00	-73.1
	3	235.2	1.65	-44.8	470.4	1.35	-91.9
	4	176.4	1.35	-43.8	352.8	1.10	-77.2

Table 6.14: System B: Maximum power gain results, $L = 32$ and $L = 64$

6.6.3 System B: Bit-flipping algorithm with PRF control

These results are shown in tables 6.14 and 6.15. In all cases, at the maximum stable input ($A_s = 0.3$), the PRF given by equation 6.22 is obtained. The dash (-) indicates the modulator is unstable for all power gains greater or equal to 1 dB (1 dB is the minimum power gain considered). These results show that the bit-flipping algorithm with PRF control alone cannot achieve the target performance for modulator orders below or equal to eight. For $L = 32$ and $L = 64$ the target PRF is achievable, but the minimum baseband noise power (for $L = 32$, $N = 8$) is only 79.3 dB. For $L = 128$, the modulator is unstable for $F_k > 4$, therefore the lowest PRF possible is 705.6 kHz. These results show that with this algorithm, it is not possible to achieve high SNRs at the same time as low PRFs.

order	F_k	PRF(kHz)	P_m	$P_b(dB)$
3	2	1411.2	2.30	-104.9
	3	940.8	1.70	-93.5
	4	705.6	1.15	-83.8
4	2	1411.2	2.20	-121.6
	3	940.8	1.30	-104.3
	4	705.6	1.00	-93.5
5	2	1411.2	2.25	-136.7
	3	940.8	1.40	-116.4
	4	705.6	-	-
6	2	1411.2	2.05	-146.8
	3	940.8	1.30	-123.1
	4	705.6	-	-
7	2	1411.2	2.00	-156.5
	3	940.8	1.30	-130.4
	4	1411.2	-	-

Table 6.15: System B: Maximum power gain results, $L = 128$

order	F_k	PRF(kHz)	System D			System C		
			S_a	P_m	$P_b(dB)$	B_q	P_m	$P_b(dB)$
4	2	352.8	1	2.35	-70.5	0.15	2.45	-66.7
	3	235.2	1	2.25	-69.0	0.20	2.30	-64.6
5	2	352.8	1	2.25	-73.3	0.15	2.45	-71.2
	3	235.2	1	2.20	-73.6	0.20	2.25	-68.2
6	2	352.8	1	2.25	-76.2	0.15	2.40	-73.8
	3	235.2	1	2.20	-76.4	0.20	2.25	-70.9
7	2	352.8	1	2.25	-77.8	0.15	2.40	-75.7
	3	235.2	1	2.20	-78.0	0.20	2.25	-72.4
8	2	352.8	1	2.25	-78.9	0.15	2.40	-77.0
	3	235.2	1	2.20	-78.8	0.20	2.25	-73.3

Table 6.16: Comparison of System C and D: $L = 32$

6.6.4 Systems C and D: Bit-flipping algorithms with Quantizer input Bound and Alternation Constraint

In this section, the performance of the bit-flipping algorithms using the techniques discussed in section 6.3.2 are evaluated, and compared to the algorithm with only the PRF control operational.

In section 6.3.2 it was shown that the quantizer input bound and stability control can restrict the minimum PRF obtainable by the algorithm. For the results presented here, the parameter B and S_a has been chosen to be high enough so that for each value of F_k, the PRF is constant with amplitude for all inputs up to $A_s = 0.3$. The procedure used to obtain these results is as follows:

For each set of parameters (i.e. L, N, F_k) a set of results containing P_b and the PRF range has been obtained for each value of B and S_a using maximal power gain filters. In the case of the quantizer input bound, B has been increased in steps of 0.05. From the set of results, the parameters which achieve the lowest P_b for constant PRF has been selected, and these results are presented in tables 6.16 to 6.18. For conciseness, only the results for PRFs around the useful range 200-750 kHz are shown.

The key features of these results are summarised below.

- For the same maximum input level and PRF, the quantizer input bound

order	F_k	PRF(kHz)	System D			System C		
			S_a	P_m	$P_b(dB)$	B_q	P_m	$P_b(dB)$
4	2	705.6	1	2.30	-97.1	0.15	2.45	-92.9
	3	470.4	1	2.20	-95.4	0.20	2.30	-90.6
	4	352.8	1	2.05	-96.4	0.30	2.20	-90.2
	5	282.2	2	1.35	-82.8	0.40	1.90	-86.1
	6	235.2	2	1.35	-83.0	0.45	1.80	-82.8
	7	201.6	2	1.35	-83.0	0.50	1.75	-82.1
5	2	705.6	1	2.15	-105.0	0.15	2.35	-101.6
	3	470.4	1	2.10	-104.8	0.20	2.15	-98.8
	4	352.8	1	2.05	-104.8	0.25	2.05	-97.6
	5	282.2	2	1.35	-87.8	0.35	1.70	-91.3
	6	235.2	2	1.35	-88.3	0.40	1.65	-88.3
	7	201.6	2	1.35	-88.6	0.50	1.65	-87.2
6	2	705.6	1	2.10	-111.7	0.15	2.25	-108.2
	3	470.4	1	2.10	-112.1	0.20	2.10	-105.4
	4	352.8	1	2.05	-111.8	0.25	2.05	-104.5
	5	282.2	2	1.35	-91.8	0.35	1.65	-95.0
	6	235.2	2	1.35	-92.3	0.40	1.60	-91.9
	7	201.6	2	1.40	-93.2	0.45	1.60	-92.0
7	2	705.6	1	2.10	-117.4	0.15	2.25	-114.5
	3	470.4	1	2.10	-118.0	0.20	2.10	-111.1
	4	352.8	1	2.10	-118.3	0.25	2.05	-110.0
	5	282.2	2	1.35	-94.6	0.35	1.65	-99.0
	6	235.2	2	1.40	-96.0	0.40	1.60	-95.6
	7	201.6	2	1.40	-96.4	0.45	1.55	-94.3
8	2	705.6	1	2.10	-122.5	0.15	2.25	-120.0
	3	470.4	1	2.10	-123.0	0.20	2.10	-116.2
	4	352.8	1	2.10	-123.4	0.30	1.85	-109.2
	5	282.2	2	1.40	-97.7	0.35	1.60	-100.1
	6	235.2	2	1.40	-98.2	0.40	1.60	-98.8
	7	201.6	2	1.40	-98.6	0.45	1.60	-99.2

Table 6.17: Comparison of System C and D: $L = 64$

			System D			System C		
order	F_k	PRF(kHz)	S_a	P_m	$P_b(dB)$	B_q	P_m	$P_b(dB)$
3	4	705.6	1	2.25	-104.7	0.30	2.55	-100.18
	5	564.5	2	1.45	-90.8	0.45	2.45	-99.4
	6	470.4	2	1.45	-90.1	0.55	2.20	-96.5
	7	403.2	2	1.35	-95.3	0.65	2.15	-95.6
	8	352.8	3	1.05	-79.8	0.75	2.00	-94.5
	9	313.6	3	1.00	-86.5	-	-	-
	10	282.2	3	1.00	-86.6	-	-	-
4	4	705.6	1	2.10	-123.2	0.25	2.15	-117.0
	5	564.5	2	1.35	-108.6	0.35	1.85	-114.0
	6	470.4	2	1.35	-109.2	0.45	1.75	-108.3
	7	403.2	2	1.35	-109.4	0.50	1.70	-107.7
	8	352.8	3	1.00	-99.1	0.60	1.60	-105.8
	9	313.6	3	1.00	-99.7	0.70	1.55	-104.8
	10	282.2	3	1.00	-100.0	-	-	-
5	4	705.6	1	2.10	-137.8	0.30	1.90	-127.7
	5	564.5	2	1.35	-119.8	0.35	1.70	-123.1
	6	470.4	2	1.35	-120.2	0.40	1.65	-119.8
	7	403.2	2	1.35	-120.6	0.45	1.60	-118.7
	8	352.8	3	1.00	-108.3	0.55	1.45	-115.2
	9	313.6	3	1.00	-108.5	0.65	1.45	-115.0
	10	282.2	3	1.00	-108.8	0.75	1.35	-112.3
6	4	705.6	1	2.05	-150.2	0.25	2.00	-141.9
	5	564.5	2	1.40	-129.4	0.35	1.55	-128.1
	6	470.4	2	1.35	-129.7	0.40	1.50	-126.2
	7	403.2	2	1.35	-130.0	0.45	1.45	-124.9
	8	352.8	3	1.00	-115.2	0.50	1.45	-125.6
	9	313.6	3	1.00	-115.6	0.65	1.35	-121.3
	10	282.2	3	1.00	-115.9	0.75	1.15	-114.4

Table 6.18: Comparison of System C and D: $L = 128$

method generally has a greater P_m than the alternation constraint method. This means that the quantizer input bound method has inherently better stability with respect to a given PRF. Both methods have higher P_m than with the PRF constraint alone and the noise power is lower.

- To maintain constant PRF with the quantizer input bound, B must increase with F_k, therefore to maintain stability P_m decreases with F_k. The combination of these increases causes P_b to increase with F_k.

- To maintain constant PRF with the alternation constraint, S_a must increase in steps with F_k:

$$\text{A=1 for } F_k\text{=2,3,4}$$
$$\text{A=2 for } F_k\text{=5,6,7}$$
$$\text{A=3 for } F_k\text{=8,9,10}$$

- Using the alternation constraint, the noise power and P_m is relatively insensitive to variations in F_k for a constant value of S_a. However, due to the change in high-pass shaping of the bit-flipping error, the noise increases sharply with S_a. Therefore a sharp increase in noise power occurs for $F_k = 5$ and $F_k = 8$. As a consequence, for $F_k = 5, 8$, the alternation constraint gives suboptimal performance and conversely, especially good performance is achieved for $F_k = 4, 7, 10$.

The relative merits of the two systems depend on the target PRF and oversampling ratio. For a target PRF of 352.8 kHz and target noise power of 108.5 dB, with $L = 64$ the value $F_k = 4$ is required. The best performance is obtained with the alternation constraint: a noise power of -111.8 dB is achieved by a sixth order modulator with PRF and alternation constraint with the parameters $F_k = 4$, $S_a = 1$. In comparison, using the quantizer input bound, a seventh order system is required, which achieves a noise performance of 110 dB.

For $L = 128$, the value $F_k = 8$ is required to obtain the target PRF and superior performance is obtained using the quantizer input bound: a noise power of -115.2 dB is achieved using a fifth order modulator. In comparison, a sixth order modulator is required with the alternation constraint.

L	order	$S_a = 1$ PRF(kHz)	P_m	$P_b(dB)$	$S_a = 2$ PRF(kHz)	P_m	$P_b(dB)$
32	4	160.2 → 170.3	2.05	-70.2	91.3 → 96.9	1.35	-58.1
	5	159.4 → 171.5	2.10	-74.1	92.9 → 98.2	1.35	-59.0
	6	160.3 → 171.9	2.10	-76.4	90.9 → 96.7	1.40	-59.4
	7	160.6 → 170.8	2.10	-77.5	91.4 → 97.0	1.40	-58.2
	8	161.0 → 173.6	2.10	-77.9	92.3 → 97.2	1.40	-44.4
64	4	319.1 → 337.6	2.05	-96.4	181.1 → 192.3	1.35	-83.31
	5	323.6 → 341.8	2.05	-105.0	184.9 → 193.8	1.35	-88.9
	6	321.0 → 343.2	2.05	-111.9	186.5 → 198.4	1.35	-92.8
	7	322.8 → 348.0	2.10	-118.4	181.8 → 193.2	1.40	-96.6
	8	323.5 → 343.5	2.10	-123.3	182.0 → 192.8	1.40	-98.8
128	3	622.6 → 668.3	2.05	-106.1	360.3 → 379.6	1.30	-96.0
	4	637.5 → 687.8	2.05	-123.3	373.9 → 399.1	1.30	-109.6
	5	647.5 → 690.9	2.05	-137.6	370.7 → 392.0	1.35	-120.7
	6	643.7 → 695.4	2.10	-150.8	375.7 → 393.4	1.35	-130.0
	7	648.1 → 698.0	2.10	-162.0	367.3 → 391.2	1.40	-138.9

Table 6.19: System E: Maximum power gain results for $S_a = 1, 2$

6.6.5 System E: Sigma-Delta Modulator with additional loop delays

In this section results are summarised for standard modulators utilising additional delays to reduce the PRF. Due to the absence of the PRF control there is no stipulation that the PRF must remain constant. The results are given in table 6.19.

These show that for two additional delays (S_a), the PRF is substantially lower than with $S_a = 1$, but the noise power is seriously degraded.

The obtained values of P_b and P_m are very close to those obtained with the alternation constraint with high values of F_k (compare, for example, the results for $L = 64$ and $S_a = 1$ with those of table 6.17 for $F_k = 4$). This is because the alternation control can be modelled as addition delays in the loop of a standard $\Sigma\Delta$ modulator (refer to section 6.4).

The most important difference between the delaying modulator and the bit-flipping algorithm with PRF and alternation constraint is that in the former, the

PRF is lower but non-constant. It will be shown in section 6.7.2 that the variation of PRF with input amplitude makes delaying modulators unsuitable for real power D-A converters with non-idealities in the output stage.

6.6.6 System F, G and H: Look-Ahead

The system performance with look-ahead is now considered. First the use of look-ahead in conjunction with the PRF control is briefly investigated (system F), then the use of the look-ahead with the PRF control and quantizer input bound (system G) and alternation constraint (system H) is investigated.

As with the alternation constraint and quantizer input bound, the look-ahead can restrict low PRFs from being obtained and cause the PRF to become signal dependent. The results presented in the following use values of LA selected high enough so the PRF is constant. The results for systems F with $L = 64$ and $L = 128$ are presented in table 6.20. The results for $L = 128$ are not considered here because it is not possible to achieve a constant PRF below 940 kHz with 1 or 2 levels of look-ahead. Higher levels of look-ahead are not considered because the system complexity becomes prohibitive.

To maintain constant PRF, the required levels of look-ahead for different values of F_k are as follows:

LA=1 for F_k=2,3

LA=2 for F_k=4,5

The noise power increases significantly with $LA = 2$ and in this respect look-ahead is fairly similar to the alternation control (system D) with the exception that the required increase for D occurs at $F_k = 5$. As a consequence, for $F_k = 4$ system D has better performance than F. To achieve the target PRF of 352.8 kHz with $L = 64$, the value $F_k = 4$ is required, therefore the alternation constraint yields superior results. Conversely, for $F_k = 5$ the LA algorithm achieves superior noise performance, indicating that the use of the LA algorithm would be desirable in a system requiring a PRF of 282.2 kHz.

The use of look-ahead algorithm in conjunction with the quantizer input bound (system G) and alternation control (system H) is summarised in tables 6.21 and 6.22.

In these systems two levels of look ahead are required to maintain constant PRF for $F_k = 3$, whereas in system F only a single level is required. This is because

L	order	F_k	PRF(kHz)	LA	P_m	$P_b(dB)$
32	4	2	352.8	1	3.30	-78.8
		3	235.2	1	2.95	-74.2
	5	2	352.8	1	2.95	-81.9
		3	235.2	1	2.75	-78.0
	6	2	352.8	1	2.95	-86.2
		3	235.2	1	2.70	-81.3
	7	2	352.8	1	2.90	-88.9
		3	235.2	1	2.65	-83.9
	8	2	352.8	1	2.80	-89.5
		3	235.2	1	2.65	-85.1
64	4	2	705.6	1	3.20	-104.6
		3	470.4	1	2.80	-99.6
		4	352.8	2	2.40	-90.8
		5	282.2	2	2.40	-90.8
	5	2	705.6	1	2.85	-112.6
		3	470.4	1	2.70	-109.4
		4	352.8	2	2.15	-97.3
		5	282.2	2	2.15	-97.7
	6	2	705.6	1	2.75	-120.7
		3	470.4	1	2.55	-116.1
		4	352.8	2	2.15	-104.6
		5	282.2	2	2.05	-103.2
	7	2	705.6	1	2.70	-127.6
		3	470.4	1	2.45	-121.6
		4	352.8	2	2.00	-107.3
		5	282.2	2	1.95	-106.7
	8	2	705.6	1	2.70	-134.5
		3	470.4	1	2.40	-126.5
		4	352.8	2	1.90	-109.1
		5	282.2	2	1.95	-111.1

Table 6.20: System F: Maximum power gain results

order	F_k	PRF(kHz)	System H				System G			
			LA	S_a	P_m	$P_b(dB)$	LA	B	P_m	$P_b(dB)$
4	2	352.8	1	1	3.20	-83.3	1	0.25	3.45	-81.9
	3	235.2	2	1	2.70	-74.2	2	0.35	2.55	-66.5
5	2	352.8	1	1	3.05	-88.1	1	0.25	3.35	-87.9
	3	235.2	2	1	2.65	-78.8	2	0.30	2.45	-70.0
6	2	352.8	1	1	2.95	-91.7	1	0.25	3.30	-92.6
	3	235.2	2	1	2.60	-82.5	2	0.35	2.40	-72.6
7	2	352.8	1	1	2.95	-94.9	1	0.25	3.25	-95.9
	3	235.2	2	1	2.60	-85.0	2	0.45	2.40	-74.7
8	2	352.8	1	1	3.05	-98.8	1	0.25	3.25	-99.2
	3	235.2	2	1	2.60	-86.8	2	0.35	2.40	-75.5

Table 6.21: System G and H for $L = 32$

order	F_k	PRF(kHz)	System H				System G			
			LA	S_a	P_m	$P_b(dB)$	LA	B	P_m	$P_b(dB)$
4	2	705.6	1	1	3.10	-109.7	1	0.25	3.45	-108.4
	3	470.4	2	1	2.55	-100.4	2	0.30	2.50	-91.9
	4	352.8	2	1	2.55	-100.1	2	0.45	2.35	-90.3
5	2	705.6	1	1	3.00	-120.4	1	0.30	3.30	-119.5
	3	470.4	2	1	2.55	-110.3	2	0.25	2.30	-99.3
	4	352.8	2	1	2.50	-110.0	2	0.50	2.15	-95.8
6	2	705.6	1	1	2.95	-129.4	1	0.25	3.20	-128.9
	3	470.4	2	1	2.50	-118.6	2	0.30	2.25	-106.4
	4	352.8	2	1	2.45	-118.0	2	0.40	2.05	-101.8
7	2	705.6	1	1	2.90	-137.1	1	0.25	3.20	-137.6
	3	470.4	2	1	2.55	-126.3	2	0.40	2.15	-110.9
	4	352.8	2	1	2.50	-125.7	2	0.50	2.00	-105.0
8	2	705.6	1	1	2.90	-144.3	1	0.25	3.10	-144.0
	3	470.4	2	1	2.55	-132.7	2	0.25	2.15	-115.8
	4	352.8	2	1	2.50	-131.9	2	0.40	2.00	-110.7

Table 6.22: System G and H for $L = 64$

the alternation constraint and quantizer input bound interfere with the look-ahead algorithm. In other words the best bits to flip using the look-ahead algorithm are different for those in the alternation algorithm. The results for system H are still superior to all the other algorithms, however.

The performance of system G is very similar to system H for $F_k = 2$. For $F_k > 2$, the performance of system G is considerably worse, as a high value of B is required to ensure that the interference is reduced and a low PRF can be maintained over a range of input amplitudes.

6.6.7 Summary of Results

The results presented so far are summarised in table 6.23 in which systems B, C, D, F, G, H are compared (system E is not compared here because the PRF is non-constant with input amplitude). In general the noise power increases in the order $B > C > D > F > H$. System G is more difficult to classify as the noise power is strongly dependent on F_k.

The performance of the various algorithms is critically dependent on F_k and therefore also the PRF of the system. In general, the noise power increases as PRF decreases, because a greater bit-flipping rate is required. The increase in noise power with F_k is small, however, for system D where S_a is constant ($F_k = 2, 3, 4$) and for system H where S_a and LA is constant ($F_k = 3, 4$). Conversely the increase in noise power with F_k is large where an increase in S_a occurs (as is the case for system D with F_k increasing from 4 to 5). Increasing LA also causes the noise power to increase suddenly, for example in systems G and H with F_k increasing from 2 to 3.

The dependency of noise power on the required values of LA and S_a (for constant PRF) means that the gain in noise power obtained by using one system over another, is dependent on the target PRF.

A consistent result, however, is that for $F_k = 2, 3, 4$ system H has the lowest noise power (PRF control with alternation constraint and look-ahead). For $F_k = 5$, this system cannot maintain constant PRF therefore system F must be used (PRF control with look-ahead) with a considerable increase in noise power.

order	F_k	$PRF(kHz)$	$P_b(dB)$					
			B	C	D	F	G	H
4	2	705.6	-96.0	-92.9	-97.1	-104.6	-108.4	-109.7
	3	470.4	-79.7	-90.6	-95.4	-99.6	-91.9	-100.4
	4	352.8	-68.4	-90.2	-96.4	-90.8	-90.3	-100.1
	5	282.2	-	-86.1	-82.8	-90.8	-119.5	-
5	2	705.6	-104.0	-101.6	-105.0	-112.6	-119.5	-120.4
	3	470.4	-84.3	-98.8	-104.8	-109.4	-99.3	-110.3
	4	352.8	-71.8	-97.6	-104.8	-97.3	95.8	-110.0
	5	282.2	-	-91.3	-87.8	-97.7	-	-
6	2	705.6	-109.4	-108.2	-111.7	-120.7	-128.9	-129.4
	3	470.4	-89.5	-105.4	-112.1	-116.1	-106.4	-118.6
	4	352.8	-74.1	-104.5	-111.8	-104.6	-101.8	-118.0
	5	282.2	-	-95.0	-91.8	-103.2	-	-
7	2	705.6	-115.1	-114.5	-117.4	-127.6	-137.6	-137.1
	3	470.4	-91.5	-111.1	-118.0	-121.6	-110.9	-126.3
	4	352.8	-73.1	-110.0	-118.3	-107.3	-105.0	-125.7
	5	282.2	-	-99.0	-94.6	-106.7	-	-
8	2	705.6	-121.1	-120.0	-122.5	-134.5	-144.0	-144.3
	3	470.4	-91.9	-116.2	-123.0	-126.5	-115.8	-132.7
	4	352.8	-77.2	-109.2	-123.4	-109.1	-110.7	-131.9
	5	282.2	-	-100.1	-97.7	-111.1	-	-

Table 6.23: Summary of Results for $L = 64$

6.7 Optimal Algorithms, System Complexity and Non-ideal Output Stages

In this section we consider which algorithms offer the best performance in relation to their system complexity. We consider also the performance of the various algorithms with non-ideal output stages.

6.7.1 System Complexity

The optimal system to use in a given application is a complex issue and depends on the performance requirements and the acceptable system complexity. It is beyond the scope of this thesis to discuss the issues of hardware implementation, however (unpublished) work by the author has shown that for a non-multiplexed architecture (i.e. each adder maps directly onto silicon), each level of look-ahead is approximately equivalent to an increase in modulator order. A conclusion based upon noise performance versus system complexity would be that, in most cases, there are no advantages to be gained from using look-ahead, as a higher order filter can be used instead. From a system perspective, however, the use of higher order filters implies a narrower transition region, which in turn requires the use of a higher order analogue reconstruction filter. Therefore, the use of look-ahead can reduce the complexity of the reconstruction filter. To complicate the issue, the transition width depends on the power gain of the NTF which, for the maximum power gain design methodology used, is related to the relative stability of the different algorithms. A system using a highly stable algorithm can accept a higher power gain filter, which has a wider transition band and therefore requires a lower complexity analogue reconstruction filter.

To illustrate these issues, we consider five ways of achieving the target specification: PRF=352.8 kHz, $P_b = -108.5\ dB$ for a maximum input of $A_s = 0.3$, shown in table 6.24. Four of the examples are chosen because they achieve constant PRF and the best possible performance (in terms of noise power) for minimum complexity with the different classes of algorithms (alternation constraint, quantizer input bound and look-ahead). The fifth example is for the $\Sigma\Delta$ system with an additional loop delay, which does not achieve constant PRF, but has low complexity. In the table, the *relative complexity* is calculated in terms of NTF order by making the approximations that each level of look-ahead is equivalent to an increase in order,

E.g.	System								P_b	R	C
	L	N	Type	F_k	B	S_a	LA	P_m	(dB)		(kHz)
1	128	5	C	8	0.55	-	0	1.45	-115.2	10	863.1
2	64	6	D	4	-	1	0	2.05	-111.8	6	700.3
3	64	8	F	4	-	-	2	1.90	-109.1	10	404.8
4	64	5	H	4	-	1	2	2.50	-110.0	7	832.9
5	64	6	E	-	-	-	-	2.05	-111.9	6	772.0

Table 6.24: Relative Complexity (R), performance and cutoff frequency (C) of five example systems

and that the other control algorithms have negligible hardware cost in comparison to the filter hardware cost. For the $L = 128$ design, a weighting factor of two is applied to reflect the doubling in sample rate. The relative complexity is therefore an approximate relative measure of adds per unit time. Example 2 and 5 have the lowest DSP complexity, since they do not use the look-ahead algorithm and have an oversampling ratio of 64. This is followed closely by example 4, which has a low order filter but uses two levels of look-ahead.

The $-3\ dB$ cutoff frequency represents the upper edge of the transition band. This is measured relative to the out of band noise level at $Lf_s/2$, which is approximately equal to $-42\ dB$ in all four examples. The wideband spectral responses of the five systems are shown in figures 6.36 to 6.38. Example 1 has the highest cutoff frequency (note the doubling of sample rate in the plot), followed closely by example 4 and 5.

For the target system parameters examples 2, 4 and 5 all have a good tradeoff between DSP complexity and analogue complexity. Example 3 has a poor tradeoff.

6.7.2 Performance with non-ideal output stage

In this section a simulation of unmatched rise and fall times in the power switch (refer to appendix E.2) is used to assess the performance of the five examples under non-ideal conditions. The mismatch is simulated by modifying the output of the modulator $y(k)$ using equation E.12 of appendix E.2 and a simulation is performed to determine the spectrum of the modified output. This is compared to the spectrum of the ideal output. A rather arbitrary mismatch of 1 nS has been chosen. To begin

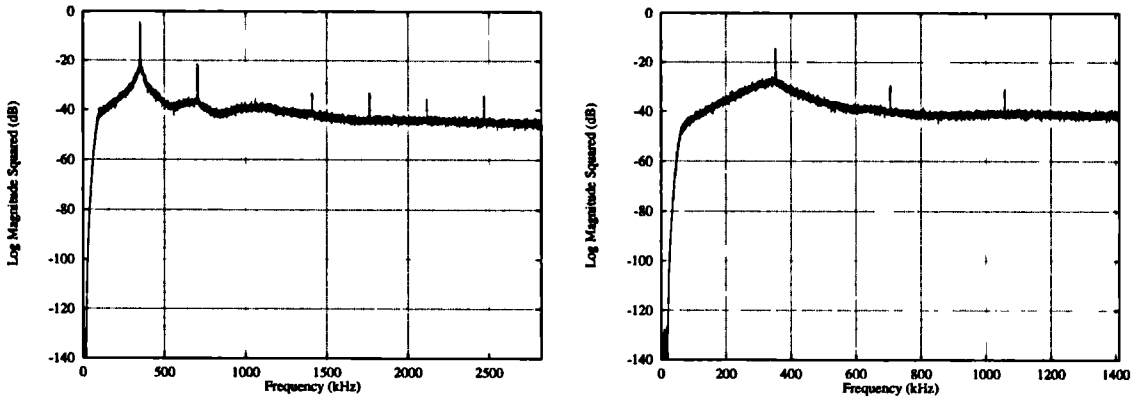


Figure 6.36: Wideband spectral response with sinewave input $A_s = 0.3$ (a) example 1, (b) example 2

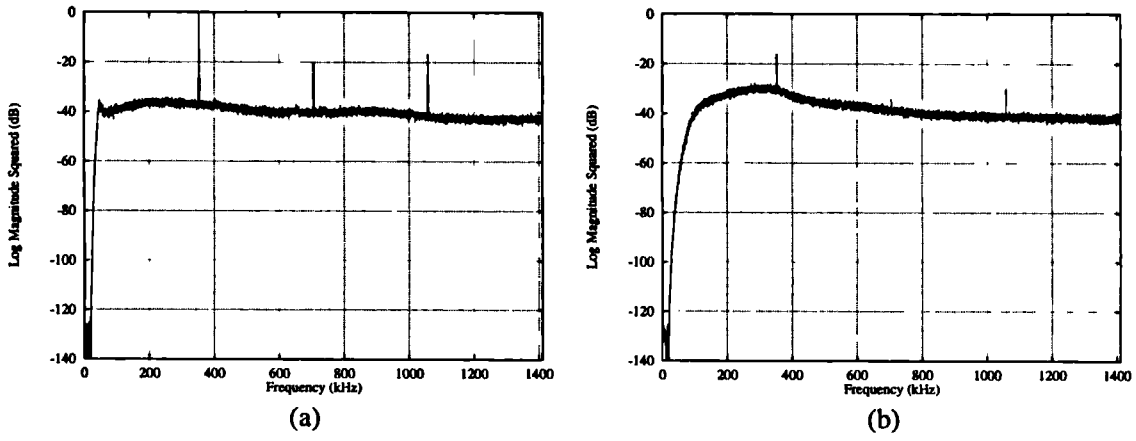


Figure 6.37: Wideband spectral response with sinewave input $A_s = 0.3$ (a) example 3, (b) example 4

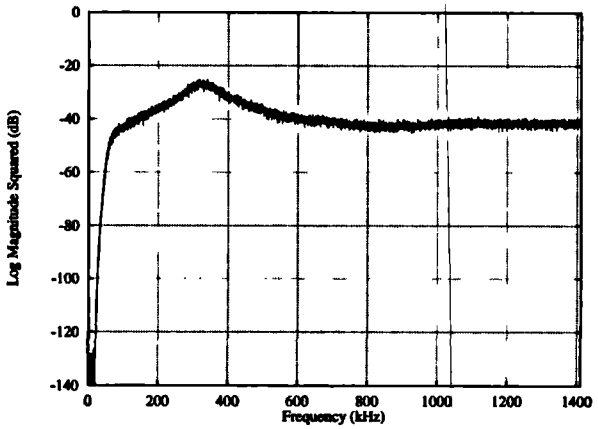


Figure 6.38: Wideband spectral response of example 5, with sinewave input $A_s = 0.3$

the investigation, results of the simulation of a standard $\Sigma\Delta$ system are given. The system parameters have been chosen to achieve the lowest possible PRF (i.e. a low oversampling, high power gain design) which meets the target noise specification: {32, 7, 4.0} (refer to table 6.12). In figure 6.39(a) the baseband spectral response with an ideal output stage (lower curve) and with a rise and fall mismatch of 1 nS (upper curve) is shown. This system achieves a noise power of -111.1 dB in the ideal case and is degraded to -83.6 dB in the non-ideal case. Second harmonic distortion is present at -90.8 dB. This distortion is predicted by the model.

The results obtained for examples 1 to 5 are shown in figures 6.39(b) to 6.41 and the results are summarised in table 6.25. In all cases the upper curves are for a rise/fall mismatch of 1 nS and the lower curves are for an ideal output stage. Especially good results are obtained for example 3. The out-of-band spectral response of this system is characterised by a high amplitude tone at the PRF and its harmonics (mainly odd order). This indicates very strong periodicity in the bitstream therefore the PRF is very uniform and there is low sensitivity to mismatched rise and fall times. The absence of peaking in the noise floor (figure 6.37(a)) suggests that the instantaneous deviation from the PRF is small. A possible reason is that this system has no alternation constraint or quantizer input bound, therefore it is usually possible to flip a bit immediately upon deviation from the average PRF (assuming that the look-ahead does not seriously constrain the bit-flipping). Conversely, when the system is controlled by the alternation constraint or quantizer input bound, the system may have to wait before bit-flipping is allowed. This would result in

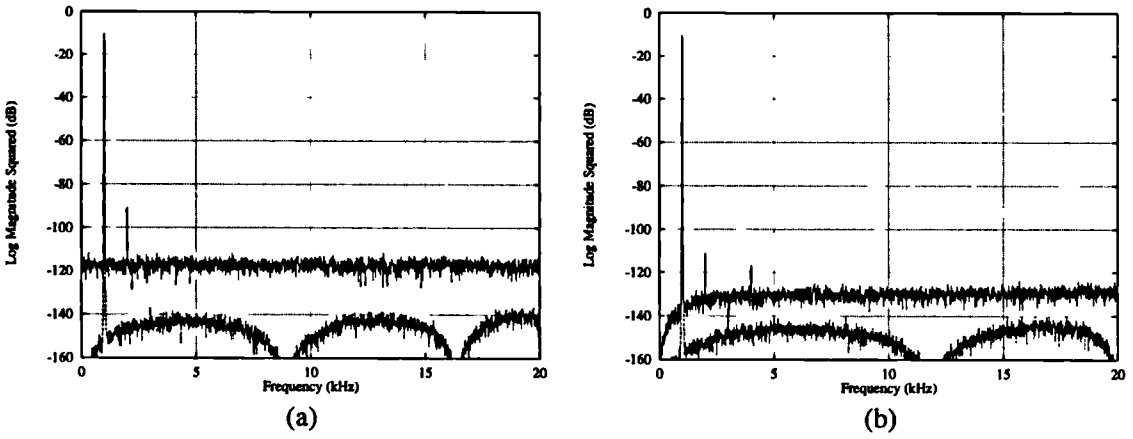


Figure 6.39: Baseband spectral responses: (a) standard system and (b) example 1

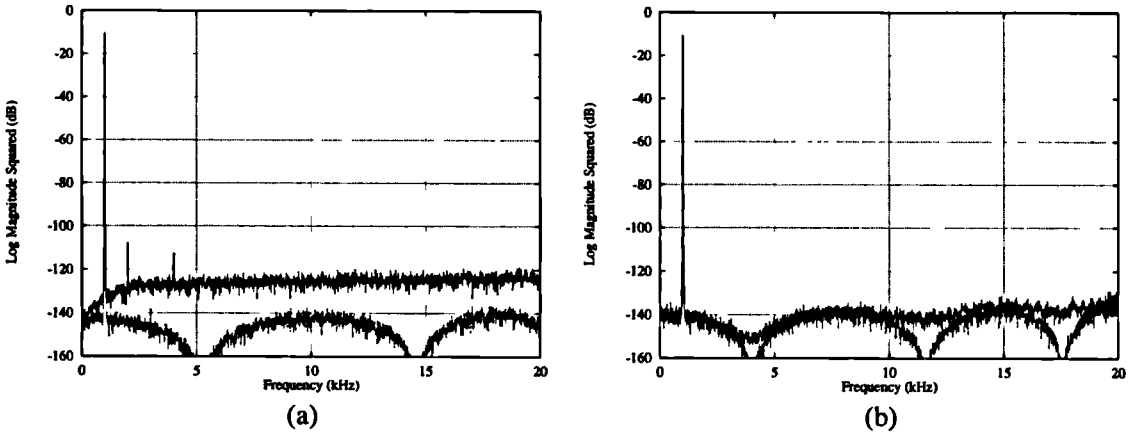


Figure 6.40: Baseband spectral responses: (a) example 2, (b) example 3

an instantaneous deviation from the average PRF, and is observed spectrally as a resonance in the noise floor around the PRF frequency and a reduction in the amplitude of the tone at the PRF. The relationship between the amplitude of the tone at the PRF and the sensitivity to mismatched rise and fall times is confirmed in table 6.25, which shows that the sensitivity to mismatch increases as the tone amplitude reduces.

The highest sensitivity to mismatch is observed in example 5 ($\Sigma\Delta$ with additional unit delay), where there is no visible tone at the average PRF, and the instantaneous deviation from the average PRF is high.

Comparing these results with table 6.24 shows that the system with highest sensitivity to a non-ideal power switch requires the greatest system complexity, and conversely the system with the highest sensitivity has the lowest complexity. This

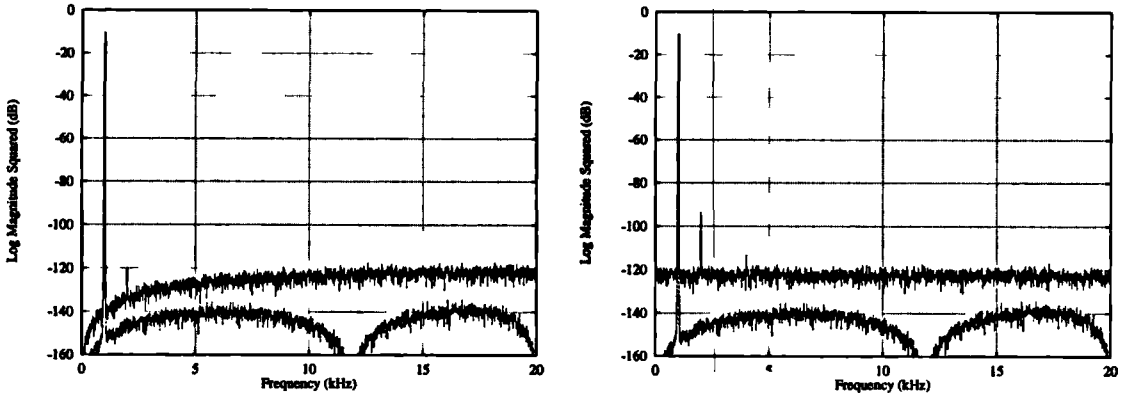


Figure 6.41: Baseband spectral responses: (a) example 4, (b) example 5

Example	P_b (ideal) (dB)	P_b (mismatched) (dB)	Max PRF tone amplitude (dB)
1	-115.2	-97.3	-4.5
2	-111.8	-92.5	-14.4
3	-109.1	-106.0	-0.07
4	-110.0	-91.3	-16.1
5	-111.9	-88.3	-

Table 6.25: Performance of example systems with mismatched rise and fall times

result is unsurprising because, in order to maintain a very constant PRF with low deviation from its average value, the degrees of freedom in the bitstream pulse locations are reduced, and so high algorithmic complexity is required to obtain acceptable stability and noise performance. All the bit-flipping algorithms have superior performance to the two $\Sigma\Delta$ systems with no bit-flipping.

6.8 Summary

Power D-A converters use a single bit D-A converter in conjunction with a power switch and passive low-pass filter to generate a high power analogue signal. High power efficiency can be obtained if the bitstream has a low average pulse repetition frequency (PRF). A PRF of around 350 kHz can result in efficiencies greater than 90 percent.

Previous research in power converters for digital audio have investigated noise-

shaped uniformly sampled pulse-width modulation (PWM) which produces a bitstream with a low PRF. PWM systems have poor linearity and require a linearisation algorithm when used for high quality audio applications. Furthermore a high system clock frequency (in the order of 90 MHz) is required to time the pulse edges and this rules out the possibility of ASIC implementation.

In this chapter, power D-A converters using $\Sigma\Delta$ modulators have been investigated. $\Sigma\Delta$ modulators have been previously ruled-out because the assumption has been made that the PRF is too high for efficient power conversion. $\Sigma\Delta$ modulators have the advantage of a low clock frequency (in the order of 3 MHz), allowing ASIC implementation. Furthermore, with an appropriate linearisation algorithm, they are capable of very high linearity. The fundamental disadvantage of $\Sigma\Delta$ modulators is the high average PRF of the bitstream.

Theoretical expressions have been derived for upper bounds on the PRF for DC and sinusoidal inputs. It has been shown that to achieve PRFs in the order of 350 kHz, an oversampling ratio below 32 and therefore an impractically high modulator order would be required. It has also been shown by simulation that the PRF is dependent on input amplitude in a nonlinear way. A model has been developed which shows that the nonlinear relationship between input amplitude and PRF causes the system to become nonlinear when mismatched rise and fall times occur in the output switch.

The use of $\Sigma\Delta$ bit-flipping algorithms has been investigated, with the aim of achieving the low PRFs of PWM conversion, combined with the linearity and low clock frequencies of $\Sigma\Delta$ modulators. The bit-flipping controls the limit cycles in the bitstream to reduce the PRF and make it constant with input amplitude. The most basic bit-flipping algorithm operates by monitoring the PRF and controlling the bit-flipping so that the PRF is reduced only when it exceeds a target PRF. The algorithm is successful as long as the PRF of the unmodified modulator does not fall below the target PRF and the modulator power gain is low enough to ensure the modulator remains stable. Using the PRF control algorithm, the modulator stability is considerably degraded and the target performance cannot be achieved. Three methods of improving the stability and noise performance have been investigated. All three achieve improved results by modifying the particular bits which are flipped.

The first uses the quantizer input bound method of chapters 4 and 5. The quantizer input bound enhances stability by ensuring that the samples which are flipped

cause a minimal increase in the quantizer error. The second method, termed the alternation constraint, reduces the baseband error of the bit-flipping, by restricting the number of allowable consecutive equi-signed bits to be flipped. The third method, termed look-ahead, predicts future samples of the quantizer output ensures that the bit-flipping only occurs if there is to be a guaranteed reduction in the PRF. This avoids unnecessary bit-flipping and enhances the stability whilst reducing the degradation in baseband noise.

The case where the alternation constraint dominates the PRF control has been analysed and found to be equivalent to a standard modulator with additional delays in the loop. A model has been developed which links the input amplitude to the PRF via the quasi-linear quantizer gain.

The performance of the systems with different oversampling ratios and target PRFs has been assessed in detail. All three systems - the alternation constraint, quantizer input bound, and look-ahead - offer advantages over the PRF control, with varying degrees of success. In general, the noise power increases with the value of the system parameters which controls the stability constraint - B, S_a, LA, therefore the choice of the control parameter is crucial to performance. Using a value which is too low, however, results in the PRF reduction being limited in a input level-dependent manner. It is therefore important that the control parameter is high enough to allow constant PRF.

For each of the algorithms there is a tradeoff between baseband noise power and PRF, with low PRFs being obtained only at the expense of high baseband noise power. The tradeoffs are different for each algorithm and can be improved by combining the algorithms. The optimal algorithm combination depends on the target performance. For the target of approximately 98 dB dynamic range and a PRF of approximately 350 kHz, the algorithm using look-ahead in conjunction with the PRF control and alternation constraint has optimal performance.

The performance of the system with a non-ideal power switch has also been considered. This study has been restricted to the case where the output switch has mismatched rise and fall times. Simulations have shown that a constant PRF which is independent of input level is desirable (as is the case with the PWM system). The choice of an optimal system entirely dependent on the accuracy of the power switch. It has been shown that systems with a high coding efficiency (i.e. the lowest noise power for a given modulator order) also have a high sensitivity to rise/fall

time mismatch. For a high accuracy power switch, the alternation constraint/look-ahead/PRF control combination is a good choice due to its high coding efficiency. For a low-accuracy power switch the look-ahead/PRF control combination is a good choice as the PRF is relatively constant.

Chapter 7

Conclusions and Further Work

A new architecture has been introduced which is applicable to $\Sigma\Delta$ modulator A-D and D-A converters, consisting of a standard modulator with the addition of a bit-flipping operator (BFO) at the output of the quantizer. The BFO selectively inverts the state of the quantizer under the control of a bit-flipping algorithm. Three distinct applications of bit-flipping have been investigated. In chapters 4 and 5, bit-flipping has been used to improve the linearity and dynamic range of low-level converters. These applications are based upon the proposition that modulators with different linearity, dynamic range, adaptive or fixed operation, differ only in their output bit-sequences. By modifying the output sequence, bit flipping potentially allows the operation and performance of the modulator to be improved. In chapter 6 bit-flipping has been used to modify the time-domain properties of the output sequence, to make $\Sigma\Delta$ modulators suitable for power D-A conversion.

The linearity achieved by $\Sigma\Delta$ modulation is damaged by the occurrence of limit cycles in the state space. The limit cycles are responsible for baseband and out-of-band tones, and noise modulation. Conventional linearisation techniques use either dither or chaos to eliminate limit cycles; however dither is inefficient to implement for A-D converters and chaos causes large increases in noise power for a given implementation complexity. Deterministic Bit-flipping (DBF) can be used to linearise the modulator by emulating dither. The technique can be modelled as a modification to the quantizer nonlinearity, which increases the complexity of the quantization process. The absence of a random component engenders the system with good relative stability and noise properties, especially where the elimination of high frequency tones is required, and makes the technique well suited to A-D conversion.

The technique is less effective at linearising third and lower order modulators and future work should concentrate on this area. One possibility is to encourage the break-up of limit cycles by combining bit-flipping with a low degree of chaos. Another is to modify the bit-flipping algorithm so that the complexity of the quantizer nonlinearity is further increased. This could be done by introducing more regions in the quantizer transfer function where bit-flipping occurs. An entirely different approach has been briefly investigated in section 4.3.1 and the paper of appendix C.1, that directly detects and breaks up limit cycles at the system output. Results have shown that certain limit cycles can be broken up by very infrequent bit-flipping, implying that very low noise penalties are possible. However, before this technique can universally applied, further work needs to be done on generalising the detection of limit cycles which are responsible for tones.

The second application of bit-flipping, which has been investigated in chapter 5 enhances the dynamic range of the modulator. The dynamic range achieved by $\Sigma\Delta$ modulation is governed by a fundamental tradeoff between stability and baseband noise power. Increasing stability by reducing the power gain of the noise transfer function (NTF) causes the transition width to decrease and therefore the baseband noise power to increase. The tradeoff can be improved by adapting the NTF so that its power gain is high at low input levels and low at high input levels. Bit flipping achieves this indirectly, by enhancing the baseband attenuation of the quantizer error spectrum under the influence of a stability condition. The bit-flipping has been shown to be equivalent to adaptively adding a low-order section to the NTF. Instead of using the input level to control the adaption, the stability of the modulator is used. If the modulator becomes unstable, the NTF adapts to reduce the power gain at the expense of an increase in baseband noise power. The key advantage is that the modulator automatically adapts to achieve the best tradeoff between noise power and stability. Accordingly, the performance of the system is critically dependent on the accuracy of the stability condition.

Future work should concentrate on improving the stability condition. One possibility is to use information from the loop-filter state-space to detect the onset of instability. More theoretical work is first required, however, to characterise the state space of higher order modulators and identify regions corresponding to saturation limit cycles. Another possibility is to use the look-ahead technique of section 6.5 to

predict the effect of bit-flipping on the future stability of the modulator. It has been shown in section 5.5 that the performance of the adaptive architecture is ultimately limited by overload in the loop, which is caused by the low-order section becoming open-loop. This could be avoided by using bit-flipping to control the NTF power gain directly by coefficient adaption, rather than by order adaption. This technique would essentially use the adaptive filter architecture of [Yu92] (described in section 2.5.1) in combination with a stability condition to control the adaption.

The third application of bit-flipping, investigated in chapter 6 modifies the time-domain properties of the bitstream — it is used to group together bits and reduce the average pulse repetition frequency (PRF) of the bitstream. This allows the $\Sigma\Delta$ modulator to be used as a viable alternative to pulse-width modulation (PWM) for power D-A conversion. In terms of the coding of information into the bitstream, the bit-flipping system can be considered to be a hybrid between pulse-width modulation (PWM) and $\Sigma\Delta$ modulation. In common with PWM it has pulses which are grouped together; however in common with $\Sigma\Delta$ modulation, there is an extra degree of freedom which allows the instantaneous PRF to vary. When combined with the use of feedback around the coding scheme, this tremendously improves its coding efficiency and consequently the bit clock frequency is considerably lower than a typical PWM system. The scheme combines the advantages of the low PRF of PWM with the low clock frequency and linearity of $\Sigma\Delta$ modulation. From an practical perspective, this opens up the possibility of ASIC implementation.

It has been shown in section 6.7.2, that when used with a non-ideal output stage, the performance of the one-bit coding scheme is degraded by increased baseband noise and distortion. The bit-flipping scheme is more sensitive to mismatched rise and fall times than PWM, but less sensitive than $\Sigma\Delta$ modulation. The sensitivity is caused by correlation between the PRF and input signal. Since this correlation is required to enhance the coding efficiency of the bitstream, it implies that high coding efficiency can only be achieved at the expense of high sensitivity to non-idealities. In order to design an optimal bit-flipping scheme, knowledge of the accuracy of the output stage is required. Therefore future research should concentrate on power switch implementation and characterisation, and develop improved output stage models to accurately predict the system performance with a given bit-flipping algorithm. This can be used to develop optimal bit-flipping algorithms. Another possible area

for future research is the use of feedback around the power switch to increase its linearity. Such a technique has been considered by [Klu92] for analogue-input $\Sigma\Delta$ power converters.

Although the above applications have used bit-flipping for different goals, it is possible to combine the goals. More specifically, in [Mag95c, Mag95b] it is shown that both the adaptive bit-flipping architectures and power D-A converter architectures have enhanced linearity over the standard modulator. The linearity of the latter architecture is very high because the high bit-flipping rate required to effect pulse grouping causes a considerable disruption to the limit cycles.

For all the above applications, significant increases in the noise and stability performance of bit-flipping modulators have been obtained by modifying the choice of bits which are flipped. Three different techniques have been investigated. The quantizer input bound of section 4.3.2, enhances stability by ensuring the quantizer error is well bounded for all samples where bit-flipping occurs. The alternation constraint of section 6.3.2 modifies the spectrum of the bit-flipping error with a high-pass characteristic to reduce the baseband noise contribution of the bit-flipping. The look-ahead technique of section 6.5 predicts the next quantizer output and uses this additional information to steer the bit-flipping algorithm. Owing to to the tradeoff between noise and stability, bit-flipping algorithms which enhance stability can be used to enhance noise performance by increasing the NTF power gain. Due to the criticality of the stability and noise performance of the bit-flipping algorithm, future studies should concentrate on finding algorithms with good performance in these aspects.

The bit-flipping algorithms have been analysed using two fundamentally different approaches. The first approach uses the PDF method of the quasi-linear model. The fundamentals of this modelling technique have been described in chapter 3. Statistical analysis is used to modelled the quantizer as a gain term followed by a additive white noise source. Information on the noise power and stability of the modulator can be gained by solving a nonlinear equation. It has been described in section 4.4.2 how bit flipping can be incorporated into the model by describing the bit-flipping operation as a modification to the quantizer nonlinearity. This is applicable if the combined quantization and bit-flipping error has a low (auto)correlation between consecutive samples, as is the case with the deterministic bit-flipping tech-

nique (DBF). A similar technique has also be used in section 4.4.1 for dithered modulators, by modelling the dither as an equivalent bit-flipping operation. The analysis reveals that DBF and dither have an almost identical affect on the noise performance of the modulator, with an increase in noise occurring as the bit-flipping rate increases. The analysis also reveals that the stability is degraded.

In the second approach, where the bit-flipping operation causes the quantizer or bit-flipping error to become auto-correlated (i.e. shaped in the frequency domain), the bit-flipping may in certain cases be mapped directly onto an equivalent operation on the NTF. Fundamentally the mapping is possible because, in the expression for the noise component for the output $E(z)NTF(z)$ (equation 2.1), a modification to $E(z)$ may be expressed instead as a modification to $NTF(z)$. This technique is applicable to weighted bit-flipping when used without the stability control (section 5.3), and also to the alternation constraint where it dominates the PRF control (section 6.4). By mapping the quantizer error shaping onto the NTF, it is then theoretically possible to use the quasi-linear model. This idea has been used in section 6.4.2 to model the alternation control, and in future work it may also be applied to weighted bit-flipping.

In addition to the theoretical studies, further work should concentrate on hardware implementation of the various bit-flipping algorithms. Work is near completion on the construction of a FPGA based 18-bit $\Sigma\Delta$ modulator incorporating the bit-flipping techniques for power D-A conversion. The prototype has verified the simulated reduction in PRF and enhanced linearity that can be obtained from the use of bit-flipping. It has also verified that bit-flipping reduces the sensitivity of the modulator to a non-ideal output stage. It is expected that the other bit-flipping algorithms will be implemented shortly, to verify the linearity of deterministic bit-flipping and fully assess the audibility of the high input-level noise power of weighted bit-flipping.

Appendix A

Simulation and Measurement

A.1 Simulation Scheme

In this appendix the simulation and measurement system which is used in all the chapters is defined. In general, the 'C' computer language has been used to implement the signal processing blocks. Double precision floating point signals have been used universally, to allow the quantization noise of the modulation process to be isolated from finite precision effects. Accordingly, no distinction is made between a D-A and A-D converter implementation. Figure A.1 shows the simulation system, which consists of the following:

- Signal Generator: DC/sinewave.
  All input signals are unquantized and generated directly at the oversampling ratio L. Therefore, no interpolator is required for the D-A converter simulation. The sampling rate of the waveform is Lf_s.

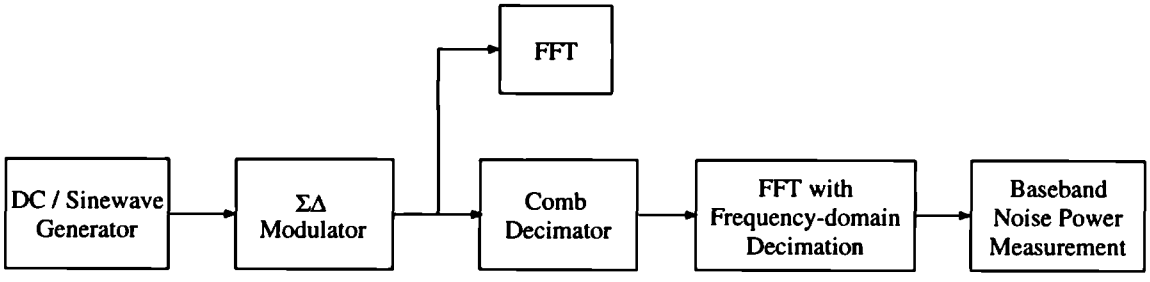


Figure A.1: Block Diagram of Simulation System

	FFT points S	Frames F	Overlap R
Baseband	8192	5	4096
Wideband	32768	30	16384

Table A.1: Simulation FFT parameters

- $\Sigma\Delta$ modulator implementing the appropriate bit-flipping architecture and associated controlling algorithm. A direct form loop filter is used with unquantized coefficients and signals. The inherent quantization effects of the simulation are below the noise floor of the modulation process. All state variables are initialised to zero.

- Decimator

  This is used for baseband spectral analysis, comprising a seven stage comb filter with decimation by a factor of eight. Integer arithmetic with wrap-around is used to eliminate integrator overflow problems [Chu84]. For flexibility, the remainder of the decimation is performed in the frequency domain. For frequency-domain decimation by a factor $D = L/8$, the baseband FFT size is multiplied by D, and only the first $1/D$ points are used in subsequent analysis.

- Fast-Fourier Transform (FFT)

  For spectral estimation, Welch's modified periodogram technique is used, which is well established and computationally efficient [Opp89]. The minimum four term Nuttall Window [Nut81] is used, which has excellent sidelobe attenuation and a small main lobe width. The FFT routine uses the NAG Fortran Library function C06FAF [Nag91]. Multiple FFT frames are used, and the squared magnitude is averaged spectrally to reduce the variance of the estimate [Opp89]. This helps to improve the discrimination between noise and periodic components. Fifty percent FFT overlapping is used so that the zero samples at the edges of one window correspond to the maximum of the next. This ensures that all the time domain samples are represented in the frequency domain. The FFT parameters used for baseband and wideband results are shown below in table A.1.

  For all simulations, an offset of 1000 samples is used in the FFT routine to reduce the measured effects of the transient response of the $\Sigma\Delta$ and decimation

filter. This means that the first 1000 samples are discarded at the input of the first FFT frame.

Ignoring the offset, the number of simulation samples generated is equal to:

$$M = L(S + (F - 1)(S - R)) \tag{A.1}$$

For baseband simulations with 64 times oversampling, approximately 1.6 million samples are processed by the modulator, corresponding to approximately half a second of audio. The aim of using such a large number of samples is to ensure that the simulation captures the behaviour of the modulator, which may change with time. In the ideal case, a larger number of simulation samples would be used, however practical considerations limit this (i.e. simulation time).

- Baseband Noise Power Measurement

The baseband noise power is the quantization noise power which falls into the baseband *referenced* to a full scale sinewave input (i.e. a sinewave of peak-to-peak amplitude equal to the quantizer step size). Noise power measurements are taken from the one-sided spectrum of the baseband FFT after compensating for the incoherent power gain of the window function [Har78]. Firstly the points containing the input signal are separated. For a signal frequency which does not correspond to an integer FFT point, spectral leakage will occur and the signal will fall into a range of points P_1 to P_2. These points must be eliminated from the noise measurement. For a S point FFT with magnitude points $|F(i)|$, the baseband noise power estimate is given by:

$$\hat{P}_b = \frac{2}{SP_w} \sum_{i=0}^{P_1-1} |F(i)|^2 + \sum_{i=P_2+1}^{(S/2)-1} |F(i)|^2 \tag{A.2}$$

This expression is the estimate of equation 2.5 and is usually expressed in deciBels. P_1 and P_2 are found in an automated procedure which distinguishes noise from signal components.

P_w is the incoherent power gain of the window function $\omega(t)$. As the window is defined in continuous time, the power gain is given by:

$$P_w = \sum_{k=0}^{S-1} \omega(kT)^2 \tag{A.3}$$

where ST is the time duration of the window.

A.1.1 Frequency Response and Power Gain Measurement

For evaluating the frequency response of the NTF, the NTF impulse response is determined and measured using an FFT with a rectangular window and a single frame of 8192 points.

It is possible to measure the power gain P_n of the NTF in the frequency domain directly from the FFT. It is more straightforward and efficient, however, to measure in the time-domain:

$$P_n = \sum_{k=0}^{M-1} ntf(k)^2 \tag{A.4}$$

P_n is expressed in deciBels in all the parameter definitions. The value of M required for accurate measurement depends on the impulse response length of the filter, which for IIR NTFs is theoretically infinite. The contributions of the impulse response to equation A.4 reduce as the oscillations decay. For practical measurement the value of M is chosen so that the last 100 contributions have a value $|ntf(k)| < 1e^{-6}$.

A.1.2 Detection of Modulator Instability

It is useful to detect the onset of $\Sigma\Delta$ modulator instability in simulations, for example in determining the maximum stable amplitude (MSA). Instability is associated with a rapid increase in the magnitude of the quantizer input amplitude, due to the presence of high-amplitude limit cycles [Hei91]. In the simulations, a modulator is deemed unstable if the quantizer input magnitude $|u(k)|$ exceeds 1000. For high order modulators this value will be rapidly reached if instability occurs.

A.2 Simulation and Modulator Parameter Definitions

The default parameters used in all the simulations are defined as follows:

- Sampling Frequency (Hz): f_s
- Oversampling Ratio: L
- Sinusoidal input signal - Peak amplitude: A_s, Frequency (Hz): f_i
- DC input signal - Amplitude: m_x
- NTF Order: N
- NTF Power gain (dB): P_n
- Oversampling rate simulation samples M

For all simulations, the sampling frequency $f_s = 44100\ Hz$ is used, making the results directly applicable to compact disk systems. The number of simulation samples M is defined according to the FFT size (equation A.1). For quasi-linear parameter evaluation (chapter 3) and pulse-repetition frequency evaluation (chapter 6), the value $M = 100000$ has been found to be sufficient for accurate results (in the sense that increasing M by orders of magnitude has very little effect on the result).

For conciseness, the following notation is also used to define the modulator NTF: $\{L, N, P_n\}$.

For example, the modulator: $\{64, 4, 4.0\}$ has a fourth order NTF with a power gain of 4.0 dB (with $K_n = 1$), designed for operation at 64 times oversampling with $f_s = 44.1\ kHz$.

A.3 Relative Tone Measurement

In this appendix we describe the measurement of tone amplitudes relative to the noise floor. This is used in chapter 4 to compare the performance of different linearisation schemes. The measurement is made relative to the noise floor because it allows a comparison of the linearity of modulators with different noise-shaping characteristics and orders, which have different levels of noise attenuation. The procedure used to evaluate the relative tone amplitude at a given frequency is to measure the largest magnitude in the FFT point associated with a particular tone frequency relative to the average magnitudes in the five FFT points either side of the tonal points (refer to figure A.2). The relative amplitude is given by equation A.5 below:

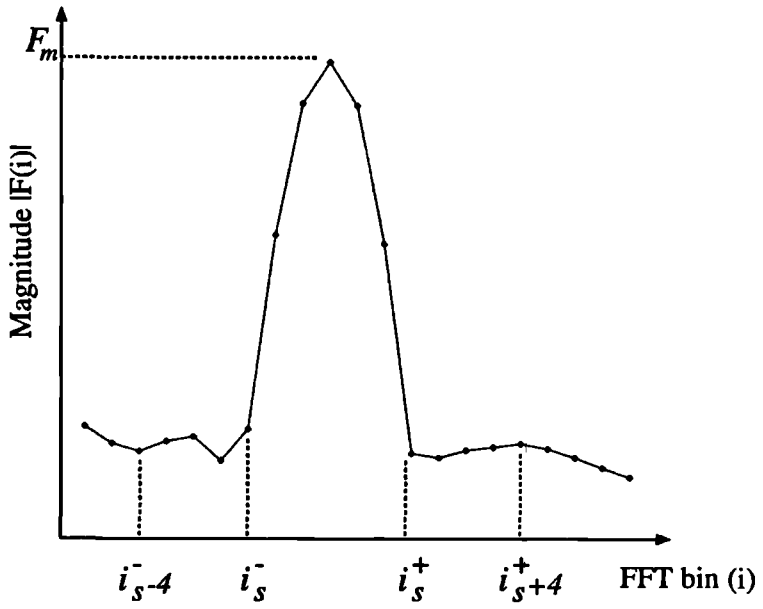


Figure A.2: Measurement of Tone Amplitudes above Noise Floor

$$A_t(dB) = 10log_{10}|F_m|^2 - 10log_{10}\frac{1}{10}\left(\sum_{i=i_s^- -4}^{i_s^-}|F(i)|^2 + \sum_{i=i_s^+}^{i_s^+ +4}|F(i)|^2\right) \qquad \text{(A.5)}$$

In this equation $|F(i)|$ is the magnitude component of point i, F_m is the maximum tone magnitude, i_s^- is the first bin to the left of the signal bins and i_s^+ is the first bin to the right of the signal bins. The FFT parameters used are given in appendix A.1.

A.4 Design Procedure for Obtaining Maximum Power Gain NTF

In this appendix, a design method is described for maximum power gain NTFs i.e. NTFs which have the largest possible power gain for stable operation with a predefined MSA. These NTFs are used in chapters 4 and 6. The NTF parameters to achieve this have been found by sweeping the NTF power gain until the required MSA is achieved. In more detail, the procedure is as follows:

1. Initialise power gain to $P_n = 5.0\ dB$.
2. Simulate modulator with appropriate bit-flipping algorithm and a 1 kHz sinewave input level $A_s = MSA$, for 1.5 million time steps.
3. If the modulator becomes unstable during the simulation, decrease P_n by 0.05 and repeat from step 2. Otherwise finish. (See appendix A.1.2 for the determination of instability).

The power gain obtained: $P_m = P_n$ represents the maximum power gain for stability with the given modulator parameters, bit-flipping scheme and MSA. P_m is also a measure of the stability of the modulator with the given bit-flipping scheme i.e. a higher P_m means the modulator has inherently higher stability margins. Note that it is not possible to directly compare stability defined by P_m *across* modulator orders without converting the power gains to the values of $min\{P_n(K_n)\}$ obtained by quasi-linear analysis (refer to section 3.5.1).

Appendix B

Appendix to Chapter 3

B.1 Accuracy of the White Noise Assumption.

In this appendix we consider the accuracy of the quantizer white noise assumption for the linear model (section 2.2.1) and quasi-linear model. The quantizer error has been determined according to figure B.1 during a simulation of the modulator. The linear model is obtained by choosing $K_n = 1$. For the quasi-linear model, K_n has been determined initially by the time-average method of section 3.3.2. A fourth order $L = 64$ modulator has been investigated with NTF power gains $P_n = 2.5, 3.0, 3.5, 4.0$ dB. The values of K_n for a DC input $m_x = 0.1$ are given in table B.1. The wideband spectra of the quantizer error for the linear and quasi-linear models are shown in figures B.2 to B.6. In all cases it can be seen that there is considerable error in the white noise assumption for the linear model. For the quasi-linear model, the noise component approximates more closely to a white noise source. The most significant error in the quasi-linear model is the presence of idle tones in the spectra. The amplitude of the tones decreases with the power gain of the NTF, therefore the white noise assumption becomes more justified for higher power gain filters.

P_n	2.0	2.5	3.0	3.5	4.0
K_n	2.94	2.34	1.93	1.66	1.45

Table B.1: Quasi-linear gains K_n for modulators plotted in figures B.2 to B.6

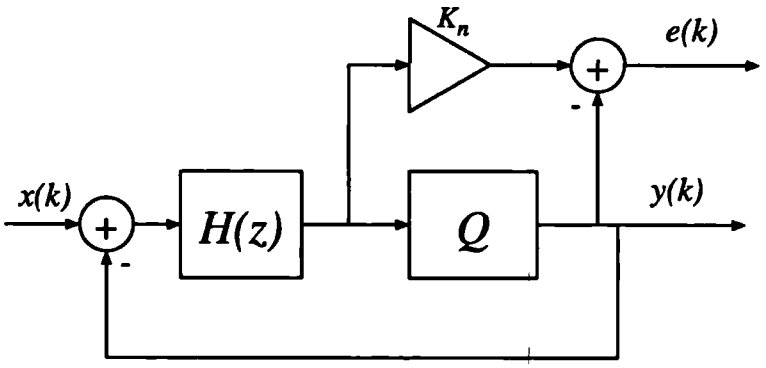


Figure B.1: Generation of Quasi-linear Error $e(k)$

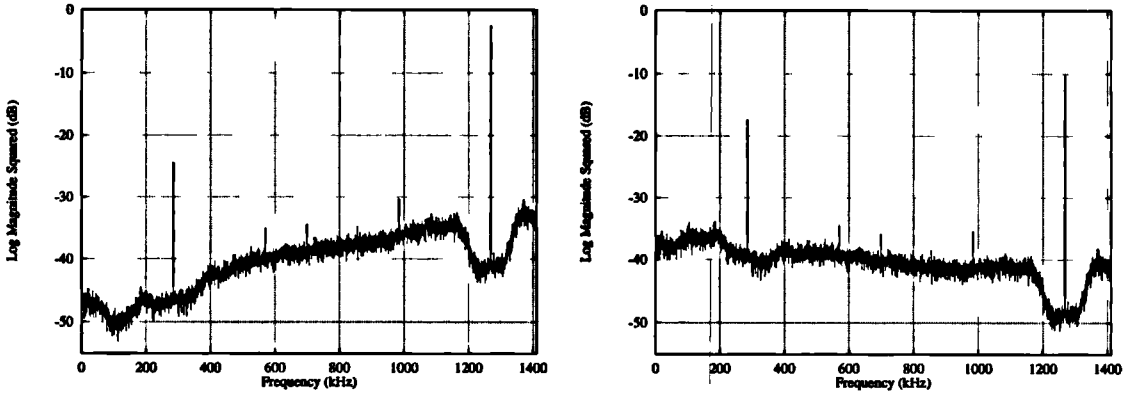


Figure B.2: Quantizer Error spectra for modulator {64, 4, 2.0} for (a) linear model and (b) quasi-linear model

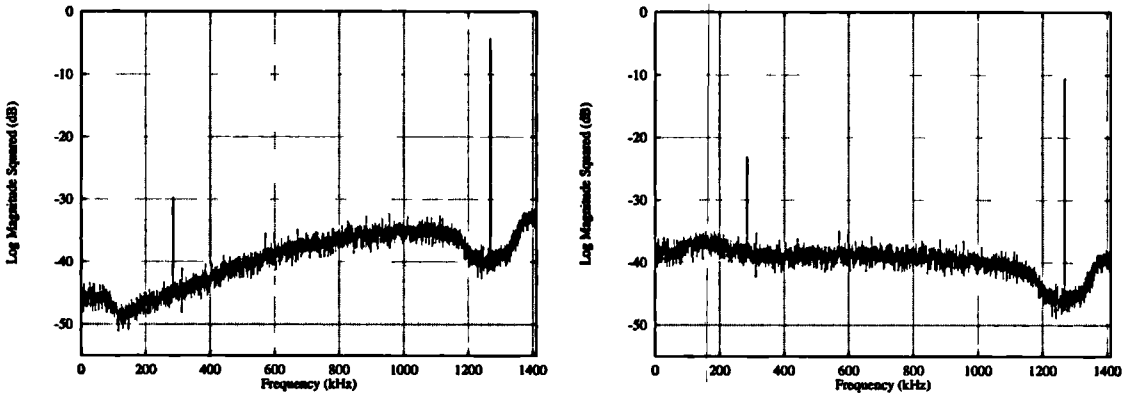


Figure B.3: Quantizer Error spectra for modulator {64, 4, 2.5} for (a) linear model and (b) quasi-linear model

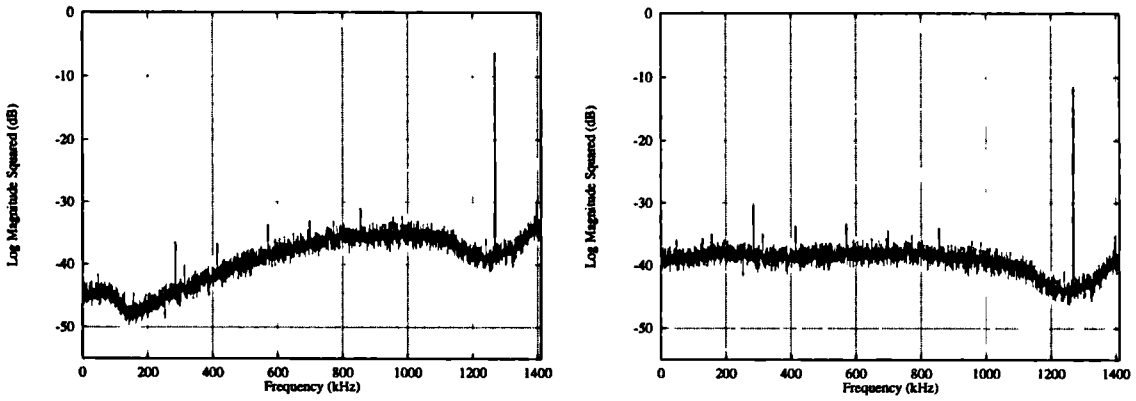


Figure B.4: Quantizer Error spectra for modulator {64, 4, 3.0} for (a) linear model and (b) quasi-linear model

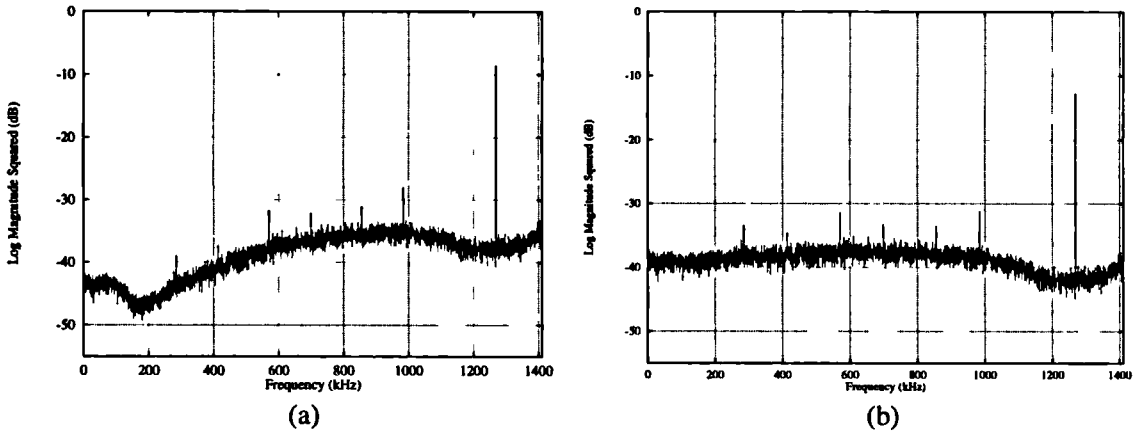


Figure B.5: Quantizer Error spectra for modulator {64, 4, 3.5} for (a) linear model and (b) quasi-linear model

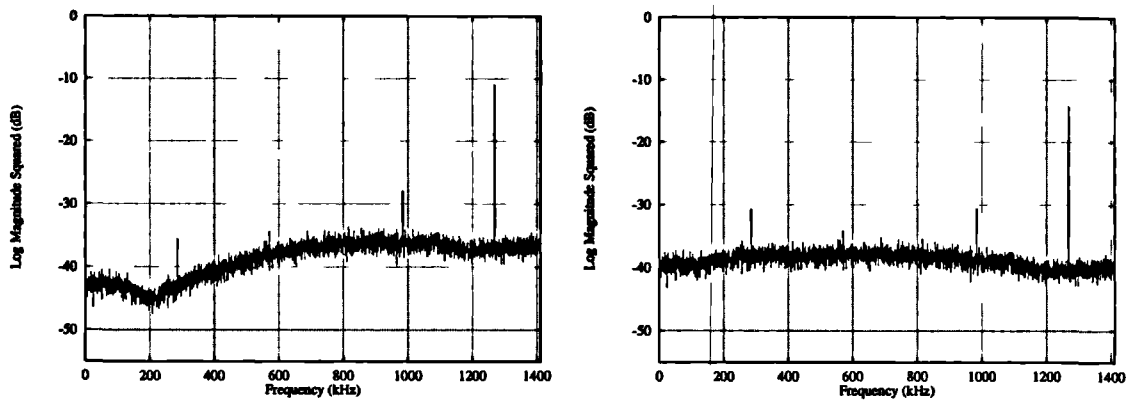


Figure B.6: Quantizer Error spectra for modulator {64, 4, 4.0} for (a) linear model and (b) quasi-linear model

B.2 Numerical Solution of a Set of Nonlinear Equations

In this appendix, the details of a nonlinear equation solver are given. This is used to find the solution to equation 3.30 for the 'standard' modulator of this chapter 4 and equation set 4.25 and 4.38 for the dithered and Deterministic Bit Flipping modulators of chapter 4. Since the latter two equations sets are in two unknowns, a general nonlinear equation solver is used: function C05NBF from the NAG Fortran library [Nag91]. This designed to solve a set of nonlinear equations in n unknowns:

$$f_i(x_1, x_2, \ldots, x_n) = 0, \qquad i = 1, 2, \ldots, n \tag{B.1}$$

The solution is obtained iteratively from an initial guess of the solution vector by minimising the sum of squares:

$$s(x) = \sum_{i=1}^{n} [f_i(x_1, x_2, \ldots, x_n)]^2 \tag{B.2}$$

using a modification of the Powell hybrid method [Nag91]. This method is insensitive to large errors in the initial guesses. Note that the two solutions to the quasi-linear gain K_n have been obtained by changing the initial guess.

For the standard modulator, the initial value of K_n for the 'real' solution has been determined using the time-average method of section 3.3.2. For most of the results, the variation in the quasi-linear parameters have been plotted for a sequence of simulations differing only by small changes in input amplitude or dither level. Due to the proximity of the solutions, only one initial guess needs to be specified by the user, and all subsequent starting points are obtained from the previous solution.

For the dithered and DBF modulators, the initial values of m_s and σ_n have been chosen heuristically.

Appendix C

Appendix to Chapter 4

C.1 Paper Reprint

This appendix is a reprint of the paper:

A.J. Magrath and M.B. Sandler.
Efficient Linearization of Sigma-Delta Modulators with Digital-Domain Dithering
Presented at the 99th Convention of the Audio Engineering Society, New York, October 1995.

The references are included in the bibliography section at the back of the dissertation.

Efficient Linearization of Sigma-Delta Modulators with Digital-Domain Dithering.

Anthony J. Magrath and Mark B. Sandler

Department of Electronic and Electrical Engineering
King's College London, Strand, London WC2R 2LS
Tel: ++44 (0)171–873 2041.
e-mail: ant@orion.eee.kcl.ac.uk

Abstract

A new technique is presented for dithering single bit sigma-delta modulator (SDM) ADCs and DACs in the digital domain. The system operates by detecting and randomizing the limit cycles responsible for baseband tones. The system is efficient to realise in hardware and avoids the implementation problems of analogue dither required in conventional SDM ADCs.

1 Introduction

One-bit Sigma-Delta Modulators (SDMs) are commonly used in high resolution analogue-to-digital converters (ADCs) and digital-to-analogue converters (DACs) for consumer and professional audio applications. Compared with traditional multibit converters they are attractive because they achieve high resolution by high precision timing rather than by precisely matched on-chip components. They can therefore be fabricated in relatively inexpensive CMOS without the need for laser trimming technology. [Nau87, Ada91, Nor89] Furthermore, SDMs can achieve resolutions greater than 16-bits and are therefore ideally suited to digital audio applications.

A well-reported disadvantage of SDMs is the presence of idle tones observable in the spectrum of the one-bit output. These tones occur several decibels above the noise floor and are related to the DC-bias in the input signal [Goud89]. For sinusoidal inputs the tones manifest themselves as harmonic and anharmonic distortion [Nau88]. They are also responsible for the modulation of the noise floor with the input signal. These effects occur predominantly at low signal levels and can therefore be objectionable to the listener.

A common remedy is the addition of random noise to the input of the quantizer in the modulator. This tends to break up the tones at the expense of a slightly increased noise floor. For ADCs the dither signal is analogue and its implementation can be problematic. Noise generators based upon thermal noise have a variance which is strongly temperature dependent and this may cause the tones suppression and stability of the modulator to become less predictable. Furthermore the noise source needs to be greatly amplified which can make power supply rejection an issue [Ris94b]. Alternatively a digital noise source can be used with a local DAC; however this approach leads to increased system complexity. Recently the use of chaotic modulators has been investigated [Sch93, Hei93b, Dun94] to alleviate the need for dither. This technique is especially attractive for ADCs, but the increased baseband noise power of chaotic modulators necessitates the use of higher order

modulators for comparable performance and this leads to increased analogue complexity. In this paper a new technique is described for dithering SDM ADCs directly in the digital domain, removing the need for an analogue dither source or the use of chaos. The technique is an extension of the method proposed in [Mag95a] and is related to the techniques discussed in [Mag95d], and leads to reduced analogue complexity. The technique is also applicable to DACs although the analogue hardware is no longer an issue because the modulator operates in the digital domain.

2 Sigma Delta Modulator Theory

The block diagram of a SDM ADC is shown in figure 1, comprising a single bit ADC (quantizer), a single bit DAC and loop filter in a feedback loop. The ADC introduces high level quantization noise which is spectrally shaped according to the loop filter to minimise the baseband content of the noise at the expense of greater noise at higher frequencies. The modulator operates on an oversampled input to give an increased bandwidth in which to shape the high frequency noise, which is later filtered by the decimation stage. The modulator operation can be described as a discrete-time system in the z-domain, in which the quantization noise is modelled as an additive error $E(z)$. For an input signal $X(z)$ and loop filter $H(z)$ the modulator output is given by

$$Y(z) = X(z)S(z) + E(z)N(z) \tag{1}$$

where

$$S(z) = \frac{H(z)}{1+H(z)} \tag{2}$$

$$N(z) = \frac{1}{1+H(z)} \tag{3}$$

The shaping of the quantization noise by the loop filter is described by the noise transfer function $N(z)$, which is usually designed as a recursive high pass filter with high attenuation in the baseband and minimal gain at higher frequencies to satisfy stability requirements [Ada91]. The assumption is often made that providing the input to the quantizer is sufficiently 'busy' and the quantizer does not overload, the quantization error is a uniformly distibuted white random process. This assumption simplifies modulator design, allowing linear design methods to be used despite the non-linearity of the quantizer in the loop [Cha90]. Studies have shown that although this is a valid approximation for multibit quantizers it is not so for single-bit quantizers and the error may be highly tonal, especially for first and second order modulators.

2.1 Mechanism of Tone Generation

The mechanism of tone generation can be understood by considering the case of a rational DC input to a modulator modelled as a discrete time system. Due to the feedback action the DC level is represented in the output by the relative occurrences of 1's and -1's. Periodic patterns in the output code are termed *limit cycles*. For zero input two limit cycles are commonly produced, which are defined here as follows:

Definition 1 *The periodic pattern {1, -1, 1, -1, 1, -1 ...} is defined as a* first order *limit cycle.*

The periodic pattern {1, 1, -1, -1, 1, 1, ...} is defined as a second order *limit cycle.*

These limit cycles are characterised spectrally by tones at $F_s/2$ and $F_s/4$ respectively, where F_s is the sampling frequency. With a small positive DC input, occasionally a '-1' is replaced by a '1' in such a way to code the correct level. The effect on the limit cycle is to reduce its fundamental frequency and produce an additional lower frequency tone. For example a DC input of 1/4 may produce a repeating limit cycle {1, -1, 1, -1, 1, -1, 1, 1} with a period of eight samples, although the exact limit cycle composition depends on the filter coefficients. In this case, the fundamental frequency is $F_s/8$ and a high frequency tone also exists (due to the remaining {-1, 1, -1, 1, 1...} pattern) at a frequency of $(1 - 1/4)(Fs/2)$ [Led3]. For a lower magnitude DC input the fundamental limit cycle period is longer and tones fall into the baseband, degrading the audio performance of the modulator. For higher order modulators the output pattern and therefore the tonal composition becomes more complex, although, despite early claims to the contrary [Cha90], for a rational input, tones still exist because the output will repeat with a finite period [Gra87].

Figure 2 shows the baseband performance of a typical undithered fourth order modulator operating at 64 times the nyquist sampling rate 44.1kHz with a DC input of 1/512. Baseband tones are present at multiples of $2822.4/512 = 5.5125\ kHz$.

Tonal artifacts are also apparent for more complex inputs; empirical evidence has shown that for sinusoidal signals harmonic distortion can be generated though often for only a limited range of input magnitudes. The sinewave signal can be perceived as a slowly varying DC input which causes the period of both the fundamental and high frequency limit cycle to vary. The limit cycle period is in effect frequency modulated by the sinewave input causing components which are related to each other by the sinewave frequency.

Another problem with SDMs is that the power of the quantization noise in the baseband varies with input signal. This *noise modulation* occurs as a result of the fact that the power of the one-bit modulator output is constant and is shared between signal, noise and distortion components [Ard87]. The presence of distortion components can lower the available noise power and since the tonal composition depends on the input signal, the noise floor varies accordingly.

All these artefacts occur especially at low signal levels and may therefore be noticeable to the listener.

3 Conventional Dithering

A common remedy is the addition of a noise source before the quantizer which randomizes the modulator limit cycles by introducing a degree of uncertainty into the modulator. Much research has been conducted into the required probability density function (PDF), amplitude and spectral density of the noise source, however all dithering systems have a common effect on the SDM, which leads to the following proposition:

Property 1 *The effect of a dither signal in a one-bit sigma-delta modulator is to change the decision made by the quantizer for appropriate samples.*

Regardless of the properties of the dither signal, the probability of the quantizer state changing depends on the instantaneous dither and filter output amplitudes. For instanta-

neous quantizer input $q_i(n)$ and dither amplitude $d(n)$ the quantizer state is only modified when

$|d(n)| > |q_i(n)|$ and $sgn|d(n)| \neq sgn|q_i(n)|$

This provides an intuitive argument to explain the observation that triangular PDF dither, with a greater probability of small amplitudes, requires a greater magnitude than a rectangular PDF dither signal to obtain the same dithering effect [Nor93]. Furthermore, the constraint indicates that a higher amplitude dither signal will cause the quantizer state to be modified for a greater proportion of samples.

4 Efficient Dithering Using Bit Flipping

This new view of dithering allows us to consider the possibility of dithering in the digital domain by inverting (or 'flipping') the state of the quantizer appropriately. The block diagram of such a system is shown in figure 3. The technique is termed *adaptive bit flipping* (ABF). The system essentially consists of a limit cycle detector (LCD) which observes the past values of the quantizer output and detects the presence of the frequently occurring first and second order limit cycles. When either limit cycle exceeds or equals a predefined length $L1$ and $L2$ samples respectively, the cycles are randomly terminated by flipping the quantizer state conditional upon $R(n) > p$. Here $0 < R(n) < 1$ is a pseudo random number sequence. In this way the limit cycle lengths are randomized. The proposed dithering technique is based on the observation that for a small input signal, an undithered modulator produces long sequences of first and/or second order limit cycles, whereas a dithered modulator produces a more random output code.

The manner in which tones are suppressed with bit-flipping can be seen by considering the example of applying a rational DC input to the system. With a DC input of 1/4 a possible repeating limit cycle is produced with a period of 8 samples such as {1 -1 1 -1 1 -1 1 1 }. The periodicity causes a tone at $Fs/8$. With the bit-flipping algorithm active and, say, $L_1 = 4$, all limit cycles equal to or longer than four samples in length are detected and randomly terminated, producing typical cycles: {1 -1 1 1 -1 1 -1 1 }, {1 -1 1 -1 -1 1 1 1} and {1 -1 1 -1 1 1 -1 1}. This assumes that the action of the feedback loop ensures that the modulator accurately codes DC despite the quantizer bit flipping. As can be seen from this example, the position of the additional '1' which codes the input is randomly modified and this influences the phase of the $Fs/8$ component, causing it to becomes more random and suppressing the tone.

Figure 4 shows the effect of dithering the modulator described in section 2.1 with ABF and parameters $L1 = L2 = 8$, $p = 0.935$.

The tone suppression technique has also been shown experimentally to apply to complex input signals.

4.1 Noise Penalty

Every quantizer bit inversion introduces an additional error in the feedback loop. The error is spectrally shaped together with the quantizer error, reducing the audio band noise contribution. In the z-domain the output of the discrete-time modulator model can be written as:

$$Q(z) = \frac{X(z)H(z)}{1+H(z)} + \frac{E(z)+F(z)}{1+H(z)} \tag{4}$$

where $X(z)$ is the input to the system, $H(z)$ is the transfer function of the loop filter, $E(z)$ and $F(z)$ are the z-transforms of the errors attributed to the quantizer error and inversion error sequence, respectively.

As with conventional dither, the additional error source increases the noise power and compromises the stability of the system [Ris94b], however at high input levels the maximum length of first and second order limit cycles tend to reduce, causing the ABF to turn off when the lengths become shorter than $L1$ and $L2$. In this way the algorithm dithers the modulator where the idle tone problem is noticeable to the listener, yet in constrast to conventional dither the maximum stable input of the modulator is not restricted. The implied noise modulation as the algorithm turns off only occurs at high input levels where signal masking considerably reduces its audibility.

4.2 High Pass Dither

There has been some interest in the use of high pass dither with SDMs, in which the dither is prefiltered by correlating dither samples in such a way to reduce the baseband amplitude of the dither signal [Nor93, Nor92, Van89]. This technique has advantages because it can reduce the noise penalty of dithered modulators. High pass dither can also be implemented with ABF by correlating samples of the bit-flipping error sequence $F(z) = Q(z) - U(z)$. This is done by ensuring that every 1 to -1 flip is followed by a -1 to 1 flip and vice-versa; this is termed an *alternation constraint*. The limit cycle detector and random condition are still present, though in general p should be reduced to compensate for the reduced probability of flipping. The alternation constraint ensures that the bit-flipping error sequence $F(z)$ has zero DC error and so reduces the in-band contribution of the dither. In figures 5 and 6 the wideband FFT of $F(z)$ is plotted for standard and high-pass ABF.

The performance of high pass ABF can be seen by considering figure 7 in which $L1 = L2 = 8$ and $p = 0.87$. These parameters are chosen so the ABF flipping activity is the same as the standard ABF scheme. In addition to a lower noise floor, the tones are further attenuated.

4.3 Noise modulation

In figure 8 the noise modulation performance of standard and high pass ABF is compared to an undithered modulator and a modulator dithered with rectangular PDF dither at a level of 1/5 quantizer step. This dither level is the minimum required to reduce noise modulation to below $0.5dB$. Each graph is generated with 1.5 million input samples for every DC increment of 0.5 dB. For the undithered modulator severe noise modulation occurs at low input levels. The variation is reduced to below 0.5 dB with rectangular PDF dither with a dynamic range penalty of 3.5 dB. Here the dynamic range is defined as the ratio of the maximum signal power and maximum noise power of the converter.

Using ABF the noise power is higher than conventional dither for low input levels but the maximum stable input is greater because the bit flipping algorithm is inactive at high input levels. The dynamic range penalties are 3.2 dB and 2.2 dB for non high-pass

and high-pass ABF, respectively. For high-pass ABF the penalty is slightly poorer than expected due to the peak in the noise floor at an input level near of 0.1 and the lower noise power at lower input levels will yield an audio advantage of a further 0.6 dB.

For the majority of the modulator range the noise modulation with ABF is below 0.5 dB. As the maximum input is approached the noise power reduces although the audibility of this noise modulation is considerably reduced due to masking by the high amplitude signal. For the example modulator the some noise modulation still occurs for particular DC inputs. This occurs because the modulator 'locks up', producing a repeating limit cycle shorter than the minimum detection lengths of the LCD. In this condition the modulator is undithered. This effect can be observed in figure 9, in which the flip rate of the modulator is plotted against DC input amplitude for a modulator using high-pass ABF. For complex inputs the deterioration in flip rate is far less evident because the modulator is kept sufficiently 'busy' to ensure that the limit cycle composition is complex.

In figure 10 the complexity of the input signal is increased by (a) the addition of a pseudo-random noise source to the DC input, at a level of $1e^{-9}$ and (b) the use of a 1 kHz sinewave test signal rather than DC. The use of a noise source reduces the noise modulation and gives some indication of the operation of the modulator with a small amount of 'analogue' noise present (note that this level of noise is insufficient to dither the modulator without bit flipping). Using a sinewave input removes the noise modulation problem altogether and suggests that a more complex music or speech signal will not cause noise modulation.

5 Design and Implementation of ABF Systems

The design of SDMs using ABF involve a compromise between stability, noise performance and efficient dithering. Figures 11 and 12 show how the tone suppression and noise power vary with the limit cycle detection lengths ($L = L1 = L2$) for high-pass ABF. Reducing p or L increases tones suppression by increasing the flip rate at the expense of noise power and ultimately stability. The choices $p = 0.935$ and $L = 8$ yield good tone attenuation without a severe noise increase. The high sensitivity of the tone suppression to small changes in the system parameters suggests that enhanced tone supression may be obtained, with only a small increase in noise penalty, by fine tuning the parameters. The detection lengths also govern the turn-off characteristics, as shown in figure 13, with shorter detection causing turn-off at higher input amplitudes. For the example modulator $L = 7$ causes the modulator to become unstable before the algorithm can turn off, whereas for $L = 10$ the algorithm turns off prematurely, reducing the signal masking of the noise modulation. A good compromise is $L = 8$.

The ABF algorithm is implemented in the digital-domain. The LCD can be implemented as a one-bit N-tap shift register, where $N = max\{L_1 - 1, L_2 - 1\}$. Combinational logic is used to detect the two limit cycles. For the detection of complete cycles ending at the current quantizer output sample Q_n, for example, the boolean expression for the LCD logic output is

$$\begin{aligned} C &= \overline{Q_n} Q_{n-1} \overline{Q_{n-2}} Q_{n-3} \cdots Q_{n-L_1-1} \\ &+ \overline{Q_n Q_{n-1}} Q_{n-2} Q_{n-3} \cdots Q_{n-L_2-1} \end{aligned}$$

Additional logic is required to detect out-of-phase limit cycles.

An efficient way of implementing the random condition is to use a technique similar to [Ada89], in which pseudo-random information in the LSBs of the decimation filter words are utilised. Here, an efficient comb filter is used in the first stage of decimation, comprising a casdade of integrators followed by sample rate reduction and a cascade of differentiators [Chu84]. The filters are implemented in 24-bit modulo (wrap-around) arithmetic. To generate a particular random condition $R(n) > p$, LSBs from the output of the cascade of integrators are ANDed together. In the system used to produce figure 14 the condition $R(n) > 0.75$ was generated by ANDing the LSB bits 1 and 2. In this case bit 0 was found to be insufficiently random to be utilised.

6 Conclusion

A dithering technique has been presented in which the limit cycles responsible for baseband tones are detected and broken up. The technique is applicable to both ADCs and DACs and the system offers hardware implementation advantages when used in ADCs because no analogue noise source (or digital noise source and DAC combination) is required. The technique reduces noise modulation at low input levels and extends the dynamic range by reverting back to an undithered modulator at high input levels.

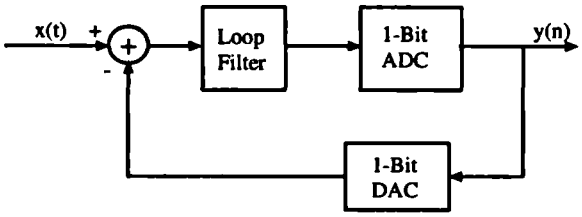


Figure 1: Block Diagram of Sigma-Delta Modulator ADC

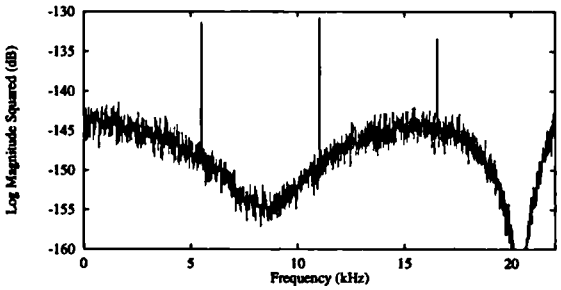


Figure 2: Baseband Spectrum of undithered SDM with 1/512 DC input

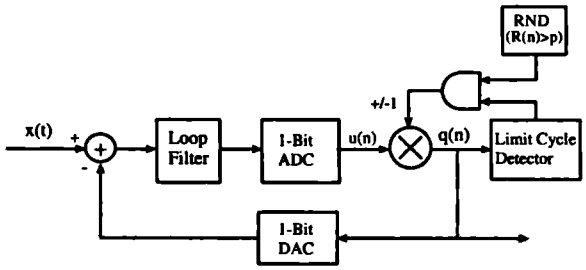


Figure 3: Block Diagram of SDM ADC with Adaptive Bit Flipping

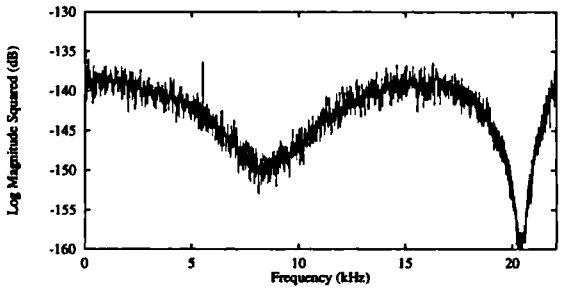


Figure 4: Baseband Spectrum of ABF with 1/512 DC input

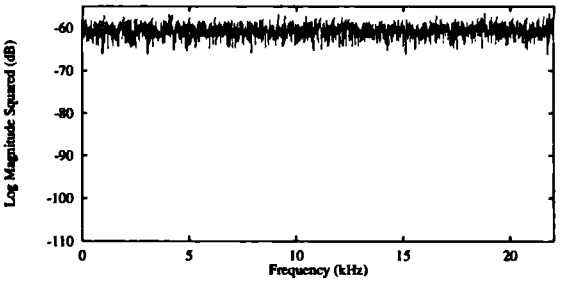


Figure 5: Baseband Spectrum of non high-pass ABF error

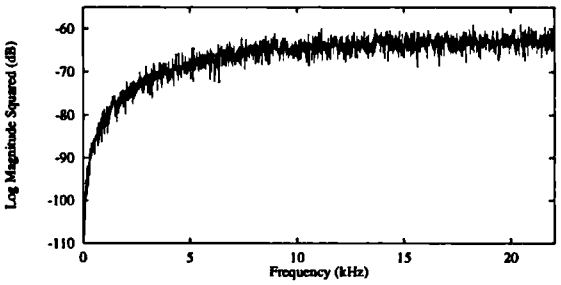


Figure 6: Baseband Spectrum of high-pass ABF error

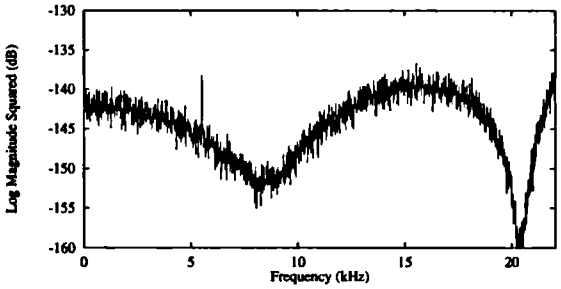


Figure 7: Baseband Spectrum of high-pass ABF with 1/512 DC input.

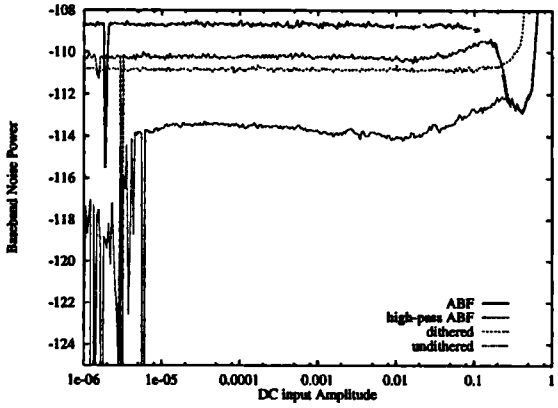


Figure 8: Noise modulation of dithering schemes

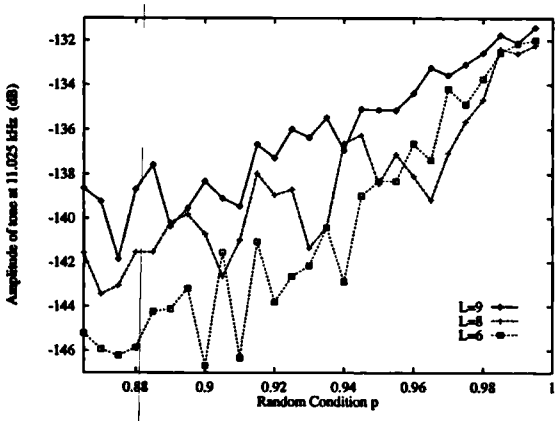


Figure 11: Amplitude of 11.025 kHz tone against Random Condition p

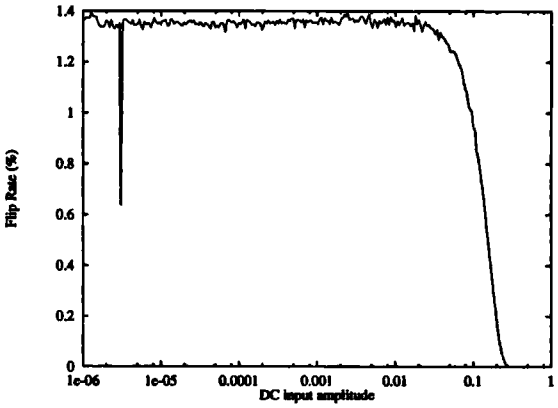


Figure 9: Flip Rate for varying DC input

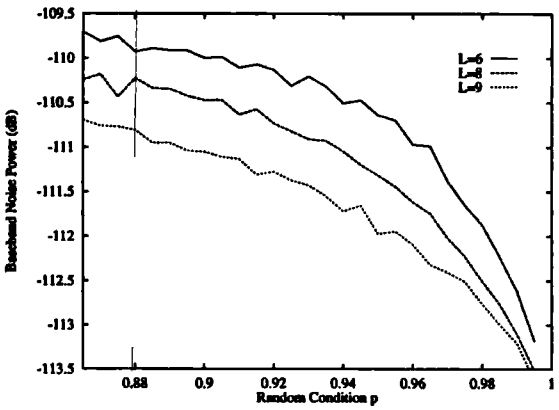


Figure 12: Baseband Noise Power of ABF system against Random Condition p

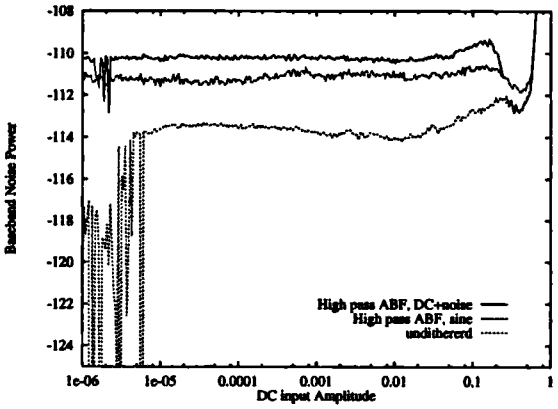


Figure 10: Noise modulation with complex input. (a) DC + Analogue noise. (b) Sine wave.

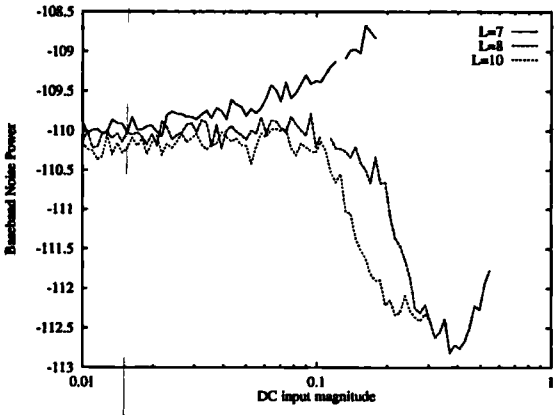


Figure 13: Baseband noise power against DC input amplitude for various L

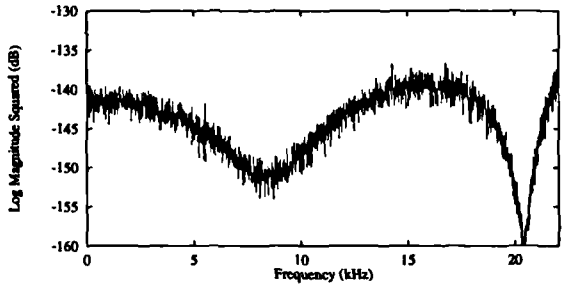


Figure 14: Baseband Spectrum of high pass ABF with 1/512 DC input and efficient implementation

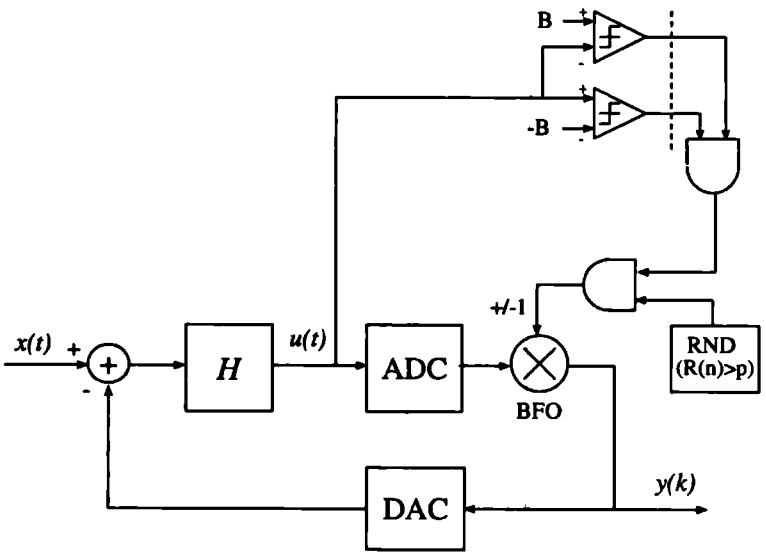


Figure C.2: A-D Converter Implementation of DBF and 1-bit dither

C.2 Implementation of Deterministic Bit-Flipping and One-bit Dither

In this appendix we consider the implementation of DBF and compare it to the implementation of one-bit dither.

It has been shown in section 4.3.4 that single bit dither is equivalent to DBF with the inclusion of a random condition 'in-line' with the BFO. Due to this equivalence, the implementation of the two schemes are similar. The implementation depends on whether the $\Sigma\Delta$ modulator is operating as an A-D or D-A converter. As an A-D, the scheme can be implemented as two comparators which detect the conditions $u < B$ and $u > -B$. The comparator outputs are clocked and ANDed in the digital domain to activate the BFO. This architecture is shown in figure C.2, in which a modification is shown to optionally implement single-bit dither by incorporating the random condition $R < 0.5$ and ANDing this with the BFO input. The random condition may be generated by means of a pseudo-random number (PRN) generator. An efficient PRN generator uses an maximum-length sequence (MLS) generator, see for example [Mut96].

The implementation for a D-A Converter involves the detection of the condition $|u(k)| < B$. A general implementation uses a magnitude detector, and comparator. The complexity of this implementation depends upon the quantization of B and

therefore the number of bits required in the comparator.

The implementation of the DBF algorithm is simpler than one-bit dither due to the elimination of the MLS generator.

C.3 Derivation of Quasi-linear Gains for Dither and Deterministic Bit-flipping

The derivations in this appendix refer to the quasi-linear analysis of section 4.4.

C.3.1 Dither

In this section the the quasi-linear gains of equations 4.16 and 4.17 are derived.

We begin with the noise gain (equation 4.14):

$$K_n = \frac{1}{\sigma_n^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{Q}\{n+r+m_s\} n p(n) p(r) dn dr \tag{C.2}$$

$p(n)$ has an assumed Gaussian distribution and $p(r)$ has a rectangular PDF of peak amplitude δ:

$$p(n) = \frac{1}{\sigma_n\sqrt{2\pi}} e^{-n^2/2\sigma_n^2} \tag{C.3}$$

$$p(r) = \begin{cases} \frac{1}{2\delta} & -\delta < r < \delta \\ 0 & \text{otherwise} \end{cases} \tag{C.4}$$

Substituting these expressions into equation C.2 we obtain:

$$K_n = \frac{1}{2\sigma_n^3\delta\sqrt{2\pi}}\int_{-\delta}^{\delta}\left\{\int_{-\infty}^{\infty}\mathcal{Q}\{n+r+m_s\} n e^{-\frac{n^2}{2\sigma_n^2}} dn\right\} dr \tag{C.5}$$

The quantizer nonlinearity of equation 4.9 is re-arranged to give:

$$\mathcal{Q}(n+r+m_s) = \begin{cases} \Delta & n \geq -(r+m_s) \\ -\Delta & n < -(r+m_s) \end{cases} \tag{C.6}$$

Leading to:

$$K_n = \frac{\Delta}{2\sigma_n^3\delta\sqrt{2\pi}}\int_{-\delta}^{\delta}\left\{-\int_{-\infty}^{-(r+m_s)} n e^{-\frac{n^2}{2\sigma_n^2}} dn + \int_{-(r+m_s)}^{\infty} n e^{-\frac{n^2}{2\sigma_n^2}} dn\right\} dr \tag{C.7}$$

$$= \frac{\Delta}{\sigma_n\delta\sqrt{2\pi}}\int_{-\delta}^{\delta} e^{-\frac{(r+m_s)^2}{2\sigma_n^2}} dr \tag{C.8}$$

The integral can be expressed in terms of the error function:

$$erf(y) = \frac{2}{\sqrt{\pi}}\int_0^y e^{-u^2} du \tag{C.9}$$

By appropriate substitution and rearrangement:

$$K_n = \frac{\Delta}{2\delta}\left\{erf\left(\frac{\delta + m_s}{\sigma_n\sqrt{2}}\right) + erf\left(\frac{\delta - m_s}{\sigma_n\sqrt{2}}\right)\right\} \tag{C.10}$$

For the signal gain we begin with equation 4.15:

$$K_s = \frac{1}{m_s}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{Q}\{n + r + m_s\}p(n)p(r)dndr \tag{C.11}$$

Substituting the dither and quantizer input PDFs:

$$K_s = \frac{1}{2\delta\sigma_n m_s\sqrt{2\pi}}\int_{\infty}^{\infty}\left\{\int_{-\delta}^{\delta}\mathcal{Q}\{n + r + m_s\}dr\right\}e^{-\frac{n^2}{2\sigma_n^2}}dn \tag{C.12}$$

We first concentrate on the evaluation of the inner integral:

$$I = \int_{\delta}^{\delta}\mathcal{Q}\{n + r + m_s\}dr \tag{C.13}$$

The result depends on the location of the step in the quantizer nonlinearity (at $r = -(n + m_s)$) in relation to δ. There are three cases:

- Case 1: $-(n + m_s) \geq \delta$

$$I = \int_{-\delta}^{\delta} -\Delta dr = -2\delta\Delta \tag{C.14}$$

- Case 2: $-(n + m_s) < -\delta$

$$I = \int_{-\delta}^{\delta} \Delta dr = 2\delta\Delta \tag{C.15}$$

- Case 3: $-\delta \leq -(n + m_s) < \delta$

$$\begin{aligned} I &= \int_{-\delta}^{-(n+m_s)} -\Delta dr + \int_{-(n+m_s)}^{\delta}\Delta dr & \text{(C.16)}\\ &= 2\Delta(n + m_s) & \text{(C.17)} \end{aligned}$$

Therefore the outer integral can be expressed as

$$\begin{aligned} K_s &= \frac{\Delta}{\delta\sigma_n m_s\sqrt{2\pi}}\left\{\int_{-\infty}^{-m_s-\delta} -\delta e^{-\frac{n^2}{2\sigma_n^2}}dn\right.\\ &+ \left.\int_{-m_s+\delta}^{\infty}\delta e^{-\frac{n^2}{2\sigma_n^2}}dn + \int_{-m_s-\delta}^{-\delta-m_s}(n + m_s)e^{-\frac{n^2}{2\sigma_n^2}}\right\}dn \end{aligned} \tag{C.18}$$

Using the error function, this expression reduces to:

$$K_s = \frac{\Delta}{2\delta m_s}\left\{(\delta+m_s)erf\left(\frac{\delta+m_s}{\sigma_n\sqrt{2}}\right)+(\delta-m_s)erf\left(\frac{\delta-m_s}{\sigma_n\sqrt{2}}\right)\right\}$$
$$+ \frac{\Delta\sigma_n}{\delta m_s\sqrt{2\pi}}\left\{e^{-\frac{(\delta+m_s)^2}{2\sigma_n^2}}-e^{-\frac{(\delta-m_s)^2}{2\sigma_n^2}}\right\} \quad \text{(C.19)}$$

C.3.2 Deterministic Bit-flipping

In this section the quasi-linear gains of equations 4.27 and 4.28 are derived.

We begin with the noise gain (equation 3.11):

$$K_n = \frac{1}{\sigma_n^2}\int_{-\infty}^{\infty} \mathcal{Q}(n+m_s)np(n)dn \quad \text{(C.20)}$$

into which the DBF quantizer nonlinearity is substituted:

$$\mathcal{Q}\{n+m_s\} = \begin{cases} \Delta & n \geq B-m_s \\ -\Delta & -m_s \leq n < B-m_s \\ \Delta & -B-m_s \leq n < -m_s \\ -\Delta & n < -B-m_s \end{cases} \quad \text{(C.21)}$$

leading to:

$$K_n = \frac{\Delta}{\sigma_n^3\sqrt{2\pi}}\left\{-\int_{-\infty}^{-B-m_s} ne^{-\frac{n^2}{2\sigma_n^2}}dn+\int_{-B-m_s}^{-m_s} ne^{-\frac{n^2}{2\sigma_n^2}}dn\right.$$
$$\left.-\int_{-m_s}^{B-m_s} ne^{-\frac{n^2}{2\sigma_n^2}}dn+\int_{B-m_s}^{\infty} ne^{-\frac{n^2}{2\sigma_n^2}}dn\right\} \quad \text{(C.22)}$$

Using the error function, this evaluates to

$$K_n = \frac{2\Delta}{\sigma_n\sqrt{2\pi}}\left\{e^{-\frac{(B+m_s)^2}{2\sigma_n^2}}+e^{-\frac{(B-m_s)^2}{2\sigma_n^2}}-e^{-\frac{m_s^2}{2\sigma_n^2}}\right\} \quad \text{(C.23)}$$

For the signal gain we begin with equation 3.12:

$$K_s = \frac{1}{m_s}\int_{-\infty}^{\infty} \mathcal{Q}(n+m_s)p(n)dn \quad \text{(C.24)}$$

Substituting the DBF quantizer nonlinearity we obtain

$$K_s = \frac{\Delta}{m_s\sigma_n\sqrt{2\pi}}\left\{-\int_{-\infty}^{-B-m_s} e^{-\frac{n^2}{2\sigma_n^2}}dn+\int_{-B-m_s}^{-m_s} e^{-\frac{n^2}{2\sigma_n^2}}dn\right.$$
$$\left.+\int_{-m_s}^{B-m_s} e^{-\frac{n^2}{2\sigma_n^2}}dn+\int_{B-m_s}^{\infty} e^{-\frac{n^2}{2\sigma_n^2}}\right\} \quad \text{(C.25)}$$

Using the error function we obtain:

$$K_s = \frac{\Delta}{m_s}\left\{erf\left(\frac{B+m_s}{\sigma_n\sqrt{2}}\right) - erf\left(\frac{B-m_s}{\sigma_n\sqrt{2}}\right) - erf\left(\frac{m_s}{\sigma_n\sqrt{2}}\right)\right\} \qquad \text{(C.26)}$$

Appendix D

Appendix to Chapter 5

D.1 Investigation of turn-on characteristics on WBF Modulators

In this appendix we investigate the behaviour of the WBF modulator with first order weighting filter, as the parameter B is increased. The analysis is based upon the dual quantizer model (figure D.1). As with the analysis of section 5.3.3, the approach taken is to use simulations to identify operating regions, then provide explanations for the existence of the regions. The operating regions identified will be shown to correspond to those of section 5.3.3, however, the analysis of the regions presented here is more detailed.

Two useful parameters are defined below:

Definition D.1 *Maximum quantizer input magnitude (MQ) is the maximum value of* $|u(k)|$.

Definition D.2 *Maximum active quantizer input magnitude (MAQ) is the maximum value of* $|u(k)|$ *for sample instances where* $y_a(k) \neq y_b(k)$.

In figures D.2 and D.3, the baseband noise power P_b is plotted against B for modulators with parameters $L = 64$, $N = 4$, power gains $P_n = 1\ dB$, $2\ dB$ and $3\ dB$ and discrete input magnitudes in the range $A_s = 0 \rightarrow 1$. Note that these results represent steady state rather than dynamic behaviour i.e. each point on the curves represents a new simulation. The noise power does not reduce smoothly with B. Instead there are three distinct regions and these are labelled in figure D.4.

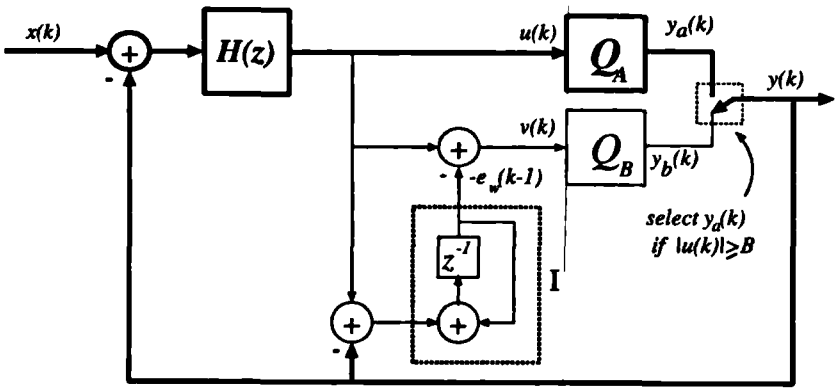


Figure D.1: Dual quantizer Model showing low-order modulator (heavy type) and internal integrator (I)

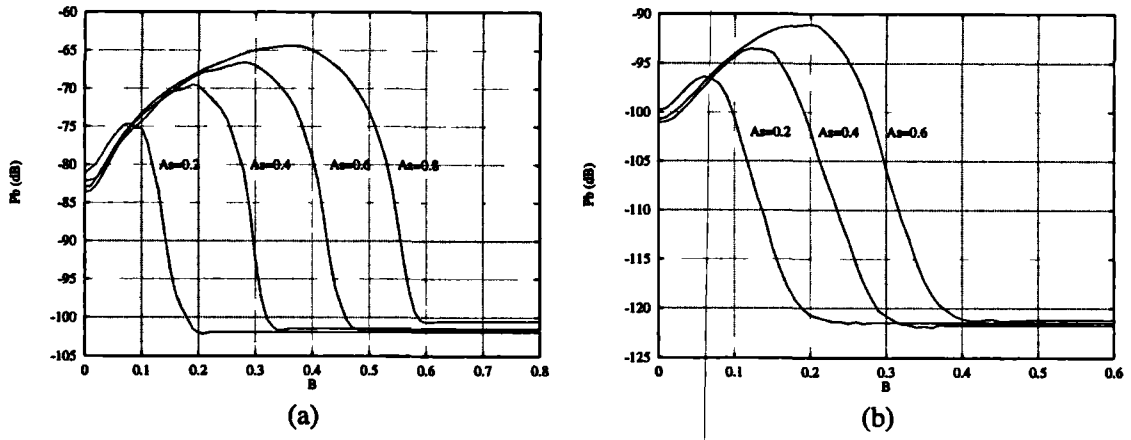


Figure D.2: Variation of Baseband noise power with B for modulators (a){64, 4, 1.0} and (b){64, 4, 2.0}

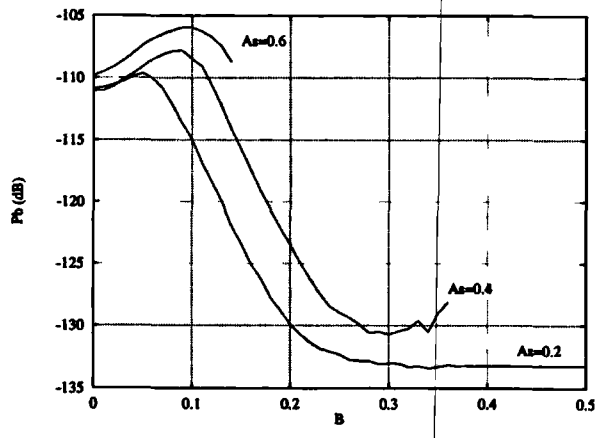


Figure D.3: Variation of Baseband noise power with B for modulator {64, 4, 3.0}

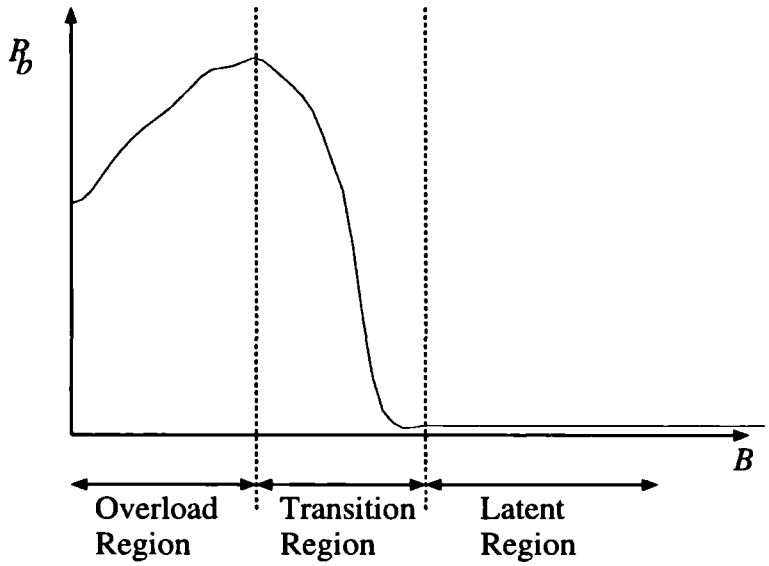


Figure D.4: Definition of Operating regions for variation in constant B

In figures D.5 and D.6 the active selection rate (ASR) is plotted against B for the same modulator parameters. The ASR is defined in section 5.3.3. It can be seen that these curves have the same general shape as the noise power curves of figures D.2 and D.3. We now consider each region of operation.

D.1.1 Overload Region

This region is characterised by the ASR and noise power increasing with B. This region is significant from a practical perspective, since it implies that using a value of B which is too low will cause the signal-to-noise ratio (SNR) to reduce rather than increase.

The case of $B = 0$ is first considered. $y(k) = y_a(k)$ for all samples and so the analysis of the system transfer functions of section 5.3.1 applies. Since quantizer Q_B is never selected, there is no feedback around the integrator I. In this condition, referring to figure 5.6 it can be seen that the quantizer input $V(z)$ represents the error signal of quantizer Q_A $(=E_a(z))$ amplified by the gain of the discrete-time integrator. Simulations have shown that $E_a(z)$ contains components of the input signal $X(z)$ due to correlation between the quantizer input and error and these components are sufficient to dominate $V(z)$. For an input sinewave of frequency f_i, The quantizer Q_B responds to the sign of $v(k)$, producing a periodic stream on average consisting of b samples of $+1$ followed by b samples of -1, where $b = LF_s/2f_i$.

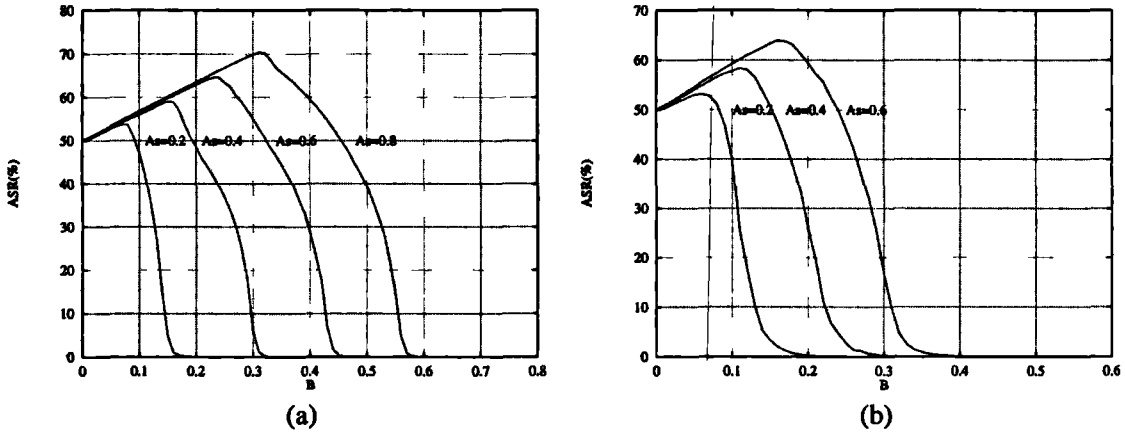


Figure D.5: Variation of ASR with B for modulators (a) {64, 4, 1.0} and (b) {64, 4, 2.0}

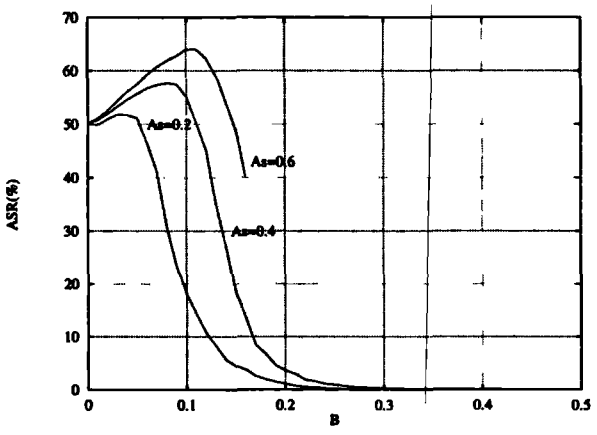


Figure D.6: Variation of ASR with B for modulator {64, 4, 3.0}

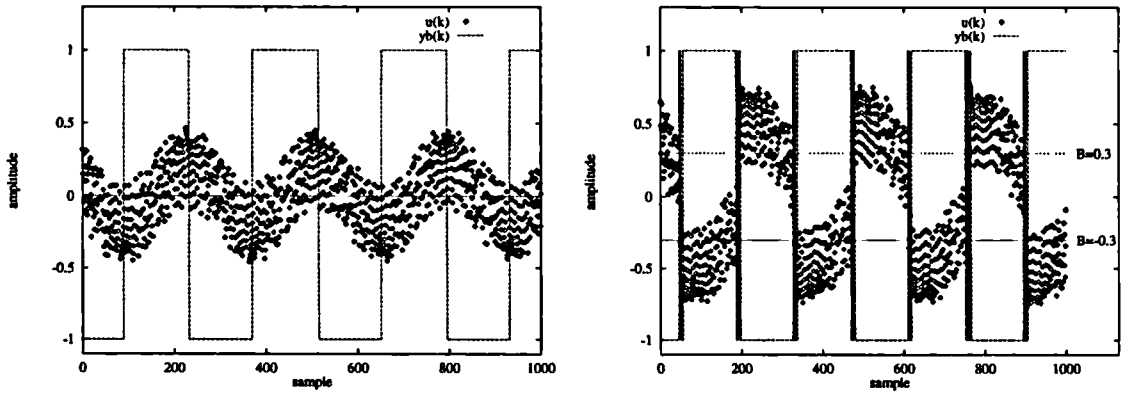


Figure D.7: Time-domain: $u(k)$ and $y_b(k)$, $P_n = 1\ dB$, $f_i = 1e4\ Hz$, $A_s = 0.8$, (a) $B = 0$ and (b) $B = 0.3$

Since $B = 0$ all samples $u(k)$ exceed B and so $ASR = 100.p\{y_a \neq y_b\}$. The values of $y_b(k)$ can be divided into two distinct regions: $y_b(k) = 1$, $y_b(k) = -1$. This is shown in the scatter graph of figure D.7(a) in which a modulator with a sinewave input of $A_s = 0.8$, $f_i = 10000Hz$ is simulated for $N = 1000$ samples. The plot shown the variation of $u(k)$ and $y_b(k)$ with sample k. In each region, due to the zero DC component of the input, $y_a(k)$ oscillates between $+1$ and -1 with equal probability, therefore, in each region $p\{y_a \neq y_b\} = 0.5$, resulting in an ASR of 50%

As B increases, the fraction of samples which meet the condition $|u(k)| > B$ reduces and it is therefore expected that the ASR will also reduce. This is clearly not observed in the previous results, which show that the ASR *increases* with B. The reason for the discrepancy will now be investigated.

Consider the region in which $y_b = -1$. As B increases, the fraction of samples α in which Q_B is selected increases. This causes an error at the input node of the system with a short-term negative DC component proportional to α. The error causes the DC component of $u(k)$ to increase and the fraction of $+1$ in the samples of y_a to increase, compensating for the fraction of samples of y_b which are selected.

In the region in which $y_b = +1$ the reverse effect occurs and the fraction of samples of y_b with value -1 increases. This effect is plotted in figure D.7(b) for a modulator with $B = 0.3$. An active quantizer selection occurs when the sign of $u(k)$ and $y_b(k)$ differ and $u(k)$ is more extreme than the horizontal lines representing $B = 0.3$. It can be seen that in each region of constant y_b, the fraction of samples for $|u(k)| \geq B$ is greater than 0.5, therefore the ASR increases.

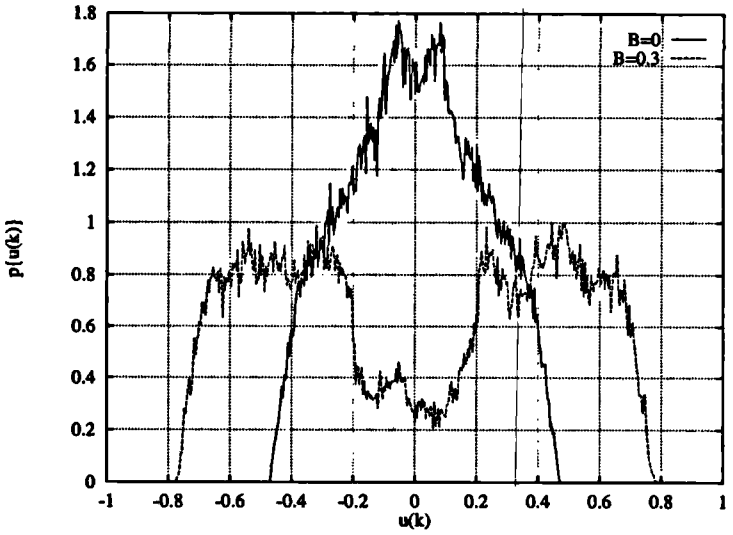


Figure D.8: PDF of $u(k)$ for $B = 0$, $B = 0.3$, $P_n = 1\ dB$, $f_i = 1e4\ Hz$, $A_s = 0.8$,

These data are also represented in the PDFs of figure D.8 for $B = 0$ and $B = 0.3$. For $B = 0.3$ the fraction of samples of $|u(k)|$ which exceed B is greater than 0.5 and since $y_a(k) \neq y_b(k)$ for these samples (figure D.7(b)), the ASR is greater than 50%.

The short term DC offsets cause the maximum value of $u(k)$ to increase and this is observed in figure D.9 in which both the MQ and MAQ are plotted against B. In the overload region MQ and MAQ are equal and increase fairly linearly with B. Equality is obtained over this region because the maximum value of $|u(k)|$ always occurs when $y_a(k) \neq y_b(k)$ (refer to the two time domain figures for an example of this). The values of MQ and MAQ are also higher for greater input amplitude, because of an increase in noise variance at the quantizer input. The increase in MQ causes the quasi-linear quantizer gain of Q_A to fall in this region and the error variance to increase. This causes the baseband noise power to increase with B.

A strange characteristic of the overload region is that y_a and y_b are fixed with opposite signs for many consecutive samples and *it is the action of the quantizer selector which codes the input signal.* For example, in the region where $y_a = +1$ the sign of $u(k)$ is positive for all samples. Quantizer Q_A is selected when $u(k) \geq B$ therefore the selector behaves as a quantizer with DC offset $+B$. In the region where $u(k)$ is negative, quantizer Q_A is selected when $u(k) < -B$, therefore the selector behaves as a quantizer with DC offset $-B$.

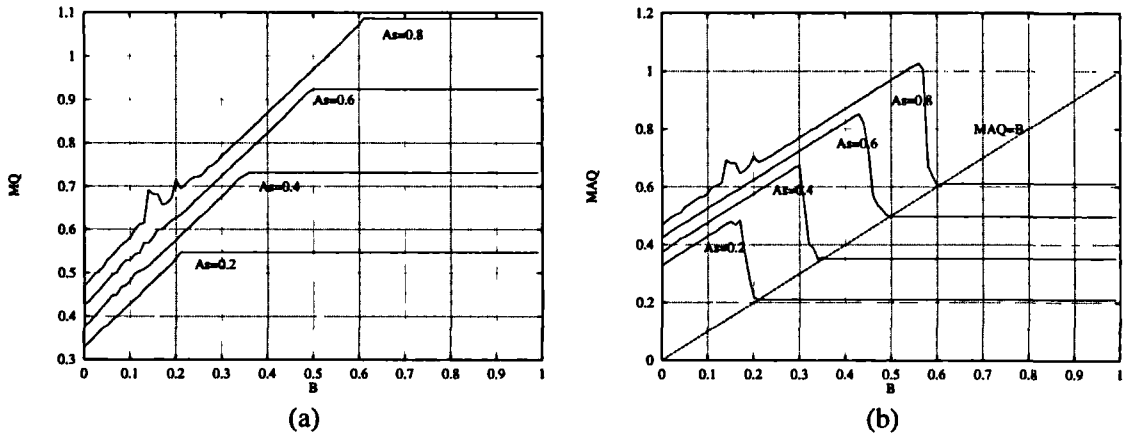


Figure D.9: Variation of MQ and MAQ with B, $P_n = 1\ dB$

D.1.2 Latent Region

This region is characterised by $ASR = 0$ and the noise power becoming constant with B.

The latent region occurs when the modulator with noise transfer function $NTF_b(z)$ is fully stable and $y_a(k) = y_b(k)$ whenever $|u(k)| \geq B$, in other words, the maximum value of $|u(k)|$ for which $y_a(k) \neq y_b(k)$ (i.e. the MAQ) is smaller than B. In the graph of MQ against B (figure D.9(b)), this region occurs to the right of the line $MAQ = B$. It can be seen that this line intersects the points corresponding to the beginning of the region in which $ASR = 0$ in figure 5.13.

The significance of the latent region is that the modulator becomes fixed and the analysis of section 5.3.1 applies, therefore the baseband noise power also remains constant with B.

D.1.3 Transition Region

This region is characterised by the ASR and noise power falling rapidly with B. The transition region represents points in the operation of the modulator between the overload and latency regions. Time-domain simulations have shown that, in this region, either mode of operation is possible, depending on the instantaneous amplitude of the input signal $x(k)$. When $x(k)$ is small, the operation is latent, and when $x(k)$ is large the operation is the same as in the overload region.

The value of B at which the transition occurs increases with input amplitude (A_s). This is because the peak amplitude of $u(k)$ increases with A_s therefore a

greater value of B is required to satisfy the condition $|u(k)| \geq B$.

Appendix E

Appendix to Chapter 6

E.1 Paper Reprint

This appendix is a reprint of the paper:

A.J. Magrath and M.B. Sandler.
Hybrid Pulse width Modulation / Sigma-Delta Modulation Power Digital-to-Analogue Converter
IEE Proceedings on Circuits, Devices and Systems, Volume 143, Number 3, June 1996.

The references are included in the bibliography section at the back of the dissertation.

Hybrid pulse width modulation / sigma-delta modulation power digital-to-analogue converter

A.J. Magrath and M.B. Sandler

Indexing Terms: DAC, Sigma-delta modulation, Pulse width modulation

Abstract

A new digital signal processing technique is presented for power digital-to-analogue converters which offers high linearity and a substantial reduction in clock frequency compared to conventional pulse-width modulation converters. The basis of the technique is to group together pulses from the output of a single-bit sigma-delta modulator. A model is derived which shows that the system output is essentially a pulse-width modulated sequence. Noise and distortion are introduced by the pulse grouping but are considerably reduced using noise-shaped feedback around the pulse grouper. Simulation results are presented which validate the model and indicate the performance of the technique with ideal and non-ideal output stages.

1

1 Introduction

In power digital-to-analogue conversion a pulse-code modulation (PCM) data stream is converted to a single bit (two-level) signal which approximates to the original signal in the low frequency portion of the spectrum. The signal controls a high power switch and the output waveform is converted to analogue by passive low-pass filtering (figure 1). The technique is especially applicable to audio applications, where the analogue amplifier can be completely removed from the audio reproduction chain, although applications also arise where precise motor control is required. High efficiencies are possible, due to the switching nature of the power stage, potentially leading to small power supplies and heatsinks.

Due to the finite switching times of practical switches, energy losses occur on every pulse transition and therefore, for the power switch to exhibit high efficiency, the average pulse repetition frequency (PRF) of the digital signal must be as low as possible. The average PRF is defined as the reciprocal of the average time between rising edges of the pulse stream. Finite switching times also influence the time-domain properties of the pulse stream and may introduce nonlinear distortion [Cra93].

Previous research has concentrated on the use of noise-shaped uniformly sampled pulse-width modulation (PWM). A b' bit PCM signal is initially oversampled L times and noise shaped to reduce the word length to b bits. The noise shaping redistributes the quantization noise to higher frequencies so that a lower wordlength is possible, at the

[1]IEE, 1996

IEE Proceedings Online no. 19960348
Paper first received 27th July 1995 and in revised form 22nd January 1996.

The authors are with the Department of Electronic and Electrical Engineering, King's College London, Strand, London WC2R 2LS, UK

expense of a higher sampling rate. The noise shaped data controls a PWM stage, where the b-bit words at sampling frequency $L.F_s$ are converted to a data stream comprising single bit words with a PRF of $L.F_s$, in which each of the possible 2^b amplitude levels corresponds to a different pulse width. Due to the discrete nature of the pulse widths it is possible to represent them in the digital domain by a sequence of single bit words at a bit rate of $2^b L.F_s$. Using this representation, it becomes apparent that a sample rate increase of 2^b occurs in the PWM stage. On a practical level, this increase causes implementation difficulties due to the high frequencies required to clock the output bits. The noise shaping considerably reduces the bit rate and the corresponding clock rate by lowering the wordlength at the input of the PWM stage without sacrificing the baseband performance of the system, however typical system parameters of $L = 8$ and $b = 8$ still yield a high bit-rate of 90.3 MHz. To achieve this bit-rate, current implementations use discrete logic counters to implement the PWM stage [Hio93b]. A lower bit-rate would be desirable to allow a more compact and cost-effective solution using, for example, ASIC technology.

A further disadvantage of PWM is that it is fundamentally non-linear, generating PRF (carrier) harmonics and sidebands, harmonic distortion of the input signal and intermodulation (foldback) noise [Pau93]. Various linearization schemes have been utilised to reduce the harmonic distortion. In [Gol93] digital signal processing is used to emulate naturally (analogue) sampled PWM, which has no harmonic distortion. In [Haw92] the non-linearity is modelled in the frequency domain and used to derive an adaptive compensation filter.

Unfortunately these schemes are unsuccessful at completely eliminating noise intermodulation, in which out-of-band noise components intermodulate with PRF and sideband components to produce lower frequency noise components which fall into the baseband [Pau93]. Both harmonic distortion *and* intermodulation noise can be reduced using feedback [Cra93], however the PWM sample rate change prevents it being applied directly and a complex architecture results.

In this paper an alternative method of generating the one-bit signal is discussed, using a modified sigma-delta modulator (SDM) architecture. SDMs achieve high linearity and relatively low bit-rates using noise-shaped feedback around a one-bit quantizer, however, the average PRF of the one-bit code is too high for efficient power DACs. A method is introduced which reduces the average PRF of a SDM bitstream. The output of the system is modelled as a PWM signal and the PWM distortion is reduced by feedback. The application of feedback is now straightforward because there is no sample rate change introduced in the conversion to PWM.

2 Noise Shaping and Sigma-Delta Modulation

The general error-feedback (noise shaper) topology of a sigma-delta modulator (SDM) is shown in figure 2. [2]. The single-bit quantizer is a two level non-linearity, which can be modelled as an additive error sequence $e(n)$. With the exception of first order SDMs, it is often assumed that the quantization error distribution is sufficiently random to be considered a noise signal [Cha90] and this assumption is especially valid when the quantizer is dithered with a random noise source. [Led3]

[2]The noise shaper topology is used throughout rather than the more common signal feedback topology because it is more suitable for the feedback structures described in section 4

In the z-domain, for an oversampled input signal $X(z)$ and feedback filter $H(z)$ the output of the system is given as:

$$V(z) = X(z) + E(z)T_N(z) \quad (1)$$
$$T_N(z) = 1 - H(z) \quad (2)$$

$T_N(z)$ is the noise-transfer function (NTF) of the system and defines the spectral characteristics of the noise in the output signal. The NTF filter is generally designed with a high pass characteristic, so that the quantization noise is attenuated in the baseband, at the expense of greater noise at higher frequencies.

For implementation, $H(z)$ must include a unit delay, which implies that the first term of the impulse response of $T_N(z)$ is unity. A general class of FIR noise-shaping filters with this property and high pass responses are given by Tewksbury in [Tew78], defined by $T_N(z) = (1 - z^{-1})^j$. These have j zeros at $z = 1$ and a $sin(\theta/2)^j$ frequency response, where θ is angular frequency ($0 < \theta < 2\pi$). SDMs using these filters are stable for $j \leq 2$ and correspond to 'standard' first and second order modulators [Can85]. For improved baseband resolution, higher order NTFs may be used, in which $T_N(z)$ is a high-pass filter designed using, e.g. a Butterworth approximation [Cha90, Ada91], however special care must be taken to ensure stability.

Predictions of noise performance for a given NTF depend on a reliable estimate of quantization noise. A general framework for this analysis is *quazi-linear modelling* [Ard87, Ris94b], in which the quantizer is modelled as a gain term K followed by a stationary white random process with variance $\sigma_n^2 = var\{e(n)\}$ (figure 3). In [Ard87] a model has been derived based on two gain terms, one for the signal and one for the noise circulating in the loop, however for noise calculations it is adequate to use only a single AC gain term [Ris94b]. For a given steady-state input signal, the value of K is a constant chosen so that the variance σ_n^2 is minimised and the error signal $e(n)$ becomes uncorrelated with the quantizer input. In [Ard87] expressions have been derived for K and σ_n^2 based upon the statistical properties of the quantizer input u(n) and output v(n) and these are repeated here:

$$K = \frac{E\{a(n)v(n)\}}{E\{a^2(n)\}} \quad (3)$$
$$\sigma_n^2 = 1 - K^2E\{a^2(n)\} - E\{v(n)\}^2 \quad (4)$$
$$a(n) = u(n) - E\{u(n)\} \quad (5)$$

where $E\{\}$ is the expectation operator. A straightforward way of evaluating K and σ_n^2 is to run a short simulation of the modulator and measure the signal expectations above. An expression for the baseband power of the quantization noise, P_b, is then found by including K into the noise transfer function:

$$T_N(z, K) = \frac{1 - H(z)}{1 + (K - 1)H(z)} \quad (6)$$
$$P_b = \frac{\sigma_n^2}{2\pi} \int_{-\frac{\pi}{L}}^{\frac{\pi}{L}} |T_N(\theta, K)|^2 \quad (7)$$

2.1 Pulse Repetition Frequency of a One-Bit Sigma-Delta Modulator.

In power DACs a low PRF is essential to ensure low power dissipation in the power switch. The PRF of the output of a SDM depends on the oversampling ratio L and the composition of the limit cycles in the output. For a stable SDM with a DC input d_i, the output signal $v(n)$ has the property $E\{v(n)\} = d_i$. For zero input level $E\{y(n)\} = 0$ and the output oscillates with limit cycles such as $1, -1, 1, -1, 1, -1, ...$ With increased input level the 'density' of 1's increases and the average PRF falls. The maximum possible PRF of a SDM is $LFs/2$, which occurs for the repeating limit cycle $1, -1, 1, -1, 1, -1, ...$ In practise this limit cycle rarely occurs, even for zero input and $1, 1, -1, -1, 1, 1, -1, ...$ patterns are more common. To achieve 16-bit resolution with a 2nd order modulator an oversampling ratio of $L = 256$ is required [Nau87]. The average PRF of this modulator with input amplitude is plotted in figure 4 which indicates a maximum PRF of 4 MHz for $F_s = 44.1kHz$. Using third or fourth order modulators, the ratio $L = 64$ is required and experiments have shown that typical modulators with this oversampling ratio reach maximum PRFs in the order of 1.2MHz, too high for efficient power switching. Furthermore, as will be seen in section 7.3 the dependence of the PRF on input amplitude can give rise to intermodulation noise when a non-ideal power switch is used. It will be seen that a *constant* PRF is desirable to reduce these effects.

3 Pulse-Group Modulation

A straightforward method of reducing the PRF of the SDM bitstream and forcing it to be constant is to group together samples with the same sign, so that after the sample-and-hold the transitions are reduced. This technique will be termed Pulse Group Modulation (PGM). The output is divided into frames of length N and the samples in each frame are relocated so that all the 1's occur in a single group. The group size is given by:

$$S(pN) = \sum_{n=0}^{N-1} v(pN - n) \tag{8}$$

where $v(n)$ are the output samples of the SDM and p is the group number.

Various grouping schemes are possible, with classifications borrowed from PWM. (figure 5). In single-sided trailing edge (SS) PGM the group begins at the start of the frame. In double-sided alternate-odd (DS) PGM an attempt is made to locate the group in the centre of the frame. For N even and S odd or vice-versa the group cannot be located at the frame centre and compensation is made by locating such groups alternately to the left then right of centre. In two sample-consecutive (TSC) PGM alternate groups are positioned at the beginning then end of the frame and this results in a reduction in PRF by a factor of two. For SS and DS PGM the PRF is LF_s/N and for TSC PGM the PRF is $LF_s/2N$.

To analyse the effect of the pulse grouping, consider a sequence, $v(n)$, of (1-bit) data from the SDM, assumed for clarity to have the values {0,1} rather than {-1, 1}. Suppose at each sample instant, the sum of the present sample and previous N-1 samples is taken:

$$y(n) = \sum_{k=0}^{N-1} v(n - k) \tag{9}$$

In the z-domain:

$$Y(z) = \sum_{k=0}^{N-1} V(z)z^{-k} = V(z)\left(\frac{1-z^{-N}}{1-z^{-1}}\right) \tag{10}$$

$$M(z) = \frac{Y(z)}{V(z)} = \frac{1-z^{-N}}{1-z^{-1}} \tag{11}$$

and so the summation is equivalent to a moving average filter (MAF) of length N . Every N^{th} sample of the summation corresponds in amplitude to the group size of each PGM pulse (figure 6). The operation of taking every N^{th} sample and discarding the remaining samples is that of decimation and the conversion to a pulse group is uniformly sampled PWM, which involves a sample rate increase by a factor N. The decimation produces aliassing and the PWM introduces harmonic distortion, carrier and sideband tones and intermodulation noise.

3.1 Levels of Aliassing Noise

The baseband noise power introduced by aliassing in the PGM system, ignoring the noise component due to the initial signal quantization is given by

$$P_a = \frac{\sigma_n^2}{2\pi N^2} \sum_{n=0}^{N-1} \int_{\frac{2\pi n}{N}-\frac{\pi}{L}}^{\frac{2\pi n}{N}+\frac{\pi}{L}} |M(\theta)|^2 |T_N(\theta, K)|^2 d\theta \tag{12}$$

The summation includes the frequency bands which alias into the baseband and the $1/N^2$ term compensates for a gain of N in the MAF. This equation predicts that, as long as the aliassing noise dominates the PWM intermodulation noise, the deterioration in signal to noise ratio will depend on the group size N, the oversampling ratio L and the SDM NTF but *not* the grouping type. The results of the numerical solution of this equation are compared to experimental results in section 7, and these results confirm the above assertion.

3.2 PWM distortion

The approximate theoretical tone spectra for uniform PWM have been derived in [Lei91] for sinewave inputs and these results are applied here to the equivalent PGM process. The tone spectrum depends on the grouping type. In the following equations D is the modulation depth ($0 < D < 1$), ω_v is the angular frequency of the input tone, $\omega_c = 2\pi LF_s/N$ is the angular frequency of the carrier, $J_n()$ is an n^{th} order Bessel function of the first kind and $\alpha = \omega_v/\omega_c$.

1. Single Side Trailing Edge PGM

$$\begin{aligned} f_{ss}(t) = & -\sum_{n=1}^{n=\infty} \frac{J_n(n\pi D\alpha)}{n\pi\alpha} \sin\left[n\omega_v t - \frac{\pi n}{2}(2\alpha+1)\right] \\ & + \sum_{m=1}^{m=\infty} \left[1+(-1)^{m+1} J_0(m\pi D)\right] \frac{\sin(m\omega_c t)}{m\pi} \end{aligned} \tag{13}$$

$$+(-1)^{(m+1)} \sum_{m=1}^{m=\infty} \sum_{n=\pm 1}^{n=\pm\infty} \frac{J_n[\pi D(m+n\alpha)]}{\pi(m+n\alpha)} \sin\left[m\omega_c t + n\omega_v t - \frac{n\pi}{2}(1+2\alpha)\right]$$

2. Double Sided Alternate odd PGM

$$\begin{aligned} f_{ds}(t) &= 2\sum_{n=1}^{n=\infty} \frac{J_n(\frac{1}{2}n\pi D\alpha)}{n\pi\alpha} \sin\left[\frac{\pi}{2}\{m+n[1-\alpha]\}\right] \cos(n\omega_v t - n\alpha\pi) \\ &+2\sum_{m=1}^{m=\infty} \frac{J_0(\frac{1}{2}m\pi D)}{m\pi} \sin(\frac{m\pi}{2}) \cos(m\omega_c t) \\ &+2\sum_{m=1}^{m=\infty} \sum_{n=\pm 1}^{n=\pm\infty} \frac{J_n[\frac{\pi D}{2}(m+n\alpha)]}{\pi(m+n\alpha)} \sin\left[\frac{\pi}{2}\{m+n(1-\alpha)\}\right] \cos\left[m\omega_c t + n\omega_v t - n\pi\alpha\right] \end{aligned} \tag{14}$$

3. Two Sample Consecutive PGM

$$\begin{aligned} f_{2sc}(t) &= 2\sum_{n=1}^{n=\infty} \frac{J_n\left(\frac{1}{2}n\pi D\alpha\right)}{n\pi\alpha} \sin\left(\frac{\pi n}{2}\right) \cos\left(n\omega_v t - \frac{n\alpha\pi}{2}\right) \\ &+2\sum_{m=1}^{m=\infty} \frac{J_0\left(\frac{1}{2}m\pi D\right)}{m\pi} \sin(\frac{m\pi}{2}) \cos(m\omega_c t) \\ &+2\sum_{m=1}^{m=\infty} \sum_{n=\pm 1}^{n=\pm\infty} \frac{J_n[\frac{\pi D}{2}(m+n\alpha)]}{\pi(m+n\alpha)} \sin\left[\frac{\pi}{2}(m+n)\right] \cos\left[m\omega_c t + n\omega_v t - \frac{n\pi\alpha}{2}\right] \end{aligned} \tag{15}$$

(16)

The spectra contain the phase distorted input frequency and harmonics, the carrier frequency and carrier harmonics with sideband terms separated by the input frequency. The harmonics and sidebands decrease monotonically with signal amplitude and also with increasing carrier (pulse repetition) frequency or reducing input frequency. For the DS and TSC modulation spectra all even harmonics of the *carrier* are zero. For TSC modulation, all even harmonics of the *input* tone are also zero and multiples of the sidebands are zero when $m+n$ is even. DS and TSC modulation types offer considerably lower levels of harmonic distortion than SS modulation [Hio93a], making them better candidates for high resolution audio.

4 Pulse Group Modulation with Feedback

To improve the distortion and noise performance of PGM, error feedback can be applied around the PGM process, taking advantage of the property that the input and output data rates are the same. This technique is equivalent to noise shaping, but here the non-linear error consists of aliassing noise, intermodulation noise and harmonic distortion, rather than quantization noise and distortion. The PGM non-linearity, which introduces the noise and distortion, is modelled in the z-domain as an additive sequence $W(z)$:

$$Y(z) = z^{-(N-1)}V(z) + W(z) \tag{17}$$

The delay occurs because the PGM block can only start outputting data when the Nth sample of the present block has been read in. The modified system structure is shown in figure 7 and the model is shown in figure 8. $E(z)$ is derived and fed back to the system input via the loop filter $G(z)$. The $z^{-(N-1)}$ delay is required to compensate for the delay in the PGM block. The noise shaper is required in the loop to requantize the input to the PGM block to one-bit. It now becomes clear the noise shaper has an advantage over the more common signal-feedback structure in that the signal transfer function is unity, which simplifies the design of $G(z)$.

The output sequence is given by

$$Y(z) = z^{-(N-1)} \{X(z) + E(z)T_N(z) + W(z)T_E(z)\} \tag{18}$$

where the quantizer noise-shaping transfer function $T_N(z)$ and the PGM error transfer function (ETF) $T_E(z)$ are given by

$$T_N(z) = 1 - H(z) \tag{19}$$
$$T_E(z) = 1 - z^{-(N-1)}G(z) \tag{20}$$

The loop filter $G(z)$ is designed so that $T_E(z)$ has a high attenuation over the baseband and so the error power in the baseband due to the pulse grouping is reduced. For implementation $G(z)$ must include a unit delay and the total z^{-N} delay prevents the system responding immediately to error disturbances. To design coefficients for $G(z)$ the class of 'Tewksbury' filters can be extended to *zero-interleaved* filters [Hio93a] by replacing z^{-1} with z^{-N} to give $T_E(z) = (1 - z^{-N})^j$. These filters are essentially comb filters with zeros at LF_s/N and the effective suppression bandwidth reduced by a factor N. The first order filter ($j = 1$) has transfer function $1 - z^{-N}$ which can be implemented by setting $G(z) = z^{-1}$. In the frequency domain.

$$\left|T_E(e^{j\theta})\right|^2 = 4\sin^2\left(\frac{\theta N}{2}\right) \tag{21}$$

The delay limits the performance of the system with feedback (section 7) and consequently it is desirable to use low values of N and obtain high SNRs at low PRFs by using low oversampling ratios and higher order loop filters instead.

5 Extension to higher order

An optimisation algorithm based on Simulated Annealing [Cat88] has been used to find suitable higher order IIR transfer functions for $H(z)$ and $G(z)$, for PGM systems used with feedback. The motivation for using optimisation is that due to the PGM delay, the first N terms in the impulse response of the ETF are fixed (equation 20) and so conventional IIR filter design methods cannot be used. The optimisation algorithm searches the multidimentional coefficient space of $H(z)$ or $G(z)$ to find coefficient sets which minimise a set of conditions or *cost functions*, in an attempt to find the most stable modulator which meets a specified baseband noise requirement.

For a general NTF, $T(z)$, the algorithm target is to minimise the infinity norm

$$|T(\theta)|_\infty^2 \text{ for } \tfrac{\pi}{L} < \theta < \pi \tag{22}$$

subject to

$$|T(\theta)|^2 < |B(\theta)|^2 \text{ for all } 0 \leq \theta \leq \frac{\pi}{L} \tag{23}$$

The constraint imposed by equation 22 restricts the maximum level of out-of-band noise produced by the noise-shaping, which is a requirement of stability in higher order modulators [Ada91]. The function $|B(\theta)|^2$ defines an upper bound on the shaping function in the baseband and to achieve a flat noise floor in the baseband, $|B(\theta)|^2$ has an inverse shape to the baseband error spectrum. The functions $T(z)$ and $|B(\theta)|^2$ are defined as follows:

1. For the SDM NTF, $T(z) = T_N(z) = 1 - H(z)$ (from equation 19). The quantizer error spectrum is assumed to be white, therefore $|B(\theta)|^2$ is a constant B_n.

2. For the PGM ETF, $T(z) = T_E(z) = 1 - z^{-(N-1)}G(z)$ (from equation 20). It can be shown that baseband power spectral density of the aliassed PGM error sequence follows a θ^2 function, and so a suitable definition of $|B(\theta)|^2$ after appropriate normalisation is:

$$|B(\theta)|^2 = B_e \left(\frac{\pi}{\theta L}\right)^2 \tag{24}$$

   where B_e is a constant

Due to interaction between the two non-linearities in the loop it is not possible to independently define target attenuations B_n and B_e for a given performance using linear analysis. Non-linear techniques are beyond the scope of this paper; however, for given system parameters N, L, a set of filters can be optimised then simulations performed to evaluate the performance of specific filter pair. An example PGM system designed using this technique is described in section 7.3.

6 Non-ideal Power Switches and Pulse Collision Detection

The analysis presented so far has assumed the output power switch is ideal i.e. the transitions occur instantaneously. In this section we consider the effects of finite rise and fall times. In terms of linear errors, finite switching times cause the magnitude response to be attenuated at high frequencies [Cra93], however we are mainly concerned here with non-linear errors i.e. how the linearity of the conversion to analogue is affected.

A simple model for the non-linear artefacts of finite switching times has been proposed in [Ris94a], in which the change in area under the output waveform is modelled in the digital domain as error impulses occurring on every transition. A simplified approach to analysing the effects of these errors is to consider how the DC component varies with the input signal level. If the PRF varies non-linearly with the input amplitude, as is the case with SDM systems (section 2.1), the DC component of the error will also vary non-linearly with the input, causing harmonic and intermodulation products to occur.

An advantage of the *PGM* system is that the PRF is generally constant giving immunity to these effects; however at very high input levels overloading may occur, causing the pulse frames to become full and adjacent pulses to 'collide'. When this occurs the

PRF decreases. A detailed discussion is not possible due to space limitations, however a solution to this problem is to introduce an additional transition by an appropriate quantizer inversion when a collision is detected. The inversion is made *inside* the noise shaping loop so that the error is compensated for in subsequent samples. For further details of related algorithms, refer to [Mag95d] and [Mag95e]. In section 7.3 the tolerance of this new scheme to unequal rise and fall times is compared to standard PGM and SDM systems.

7 Experimental and Theoretical Results

In sections 7.1 and 7.2 a system with an ideal power switch is considered. For the majority of the examples a second order modulator is used to demonstrate the technique and simplify analysis. In section 7.3 high order systems are investigated, with performance targetted at audio quality conversion, and the effects of a non-ideal output switch are also considered.

7.1 Open Loop PGM

The performance of open-loop PGM system has been evaluated by computer simulation using a 2nd order SDM dithered with rectangular PDF white noise spanning the quantizer step. The dither is used to reduce the level of idle tones present in the one-bit output of the SDM [Led3], which may alias back into the baseband with pulse-grouping.

In figure 9, a $-6dB$ 5 kHz sinewave is applied to a SS PGM system with parameters $L = 64, N = 8$. 2nd harmonic distortion and increased baseband noise are observable due to PWM and aliassing. The lower plot also shows the direct output of the SDM. Here third harmonic distortion is present due to overloading in the SDM.

The experimentally measured levels of noise introduced for SS PGM for various L and N are plotted in figure 10. On the same graph the numerical solution of equation 12 is plotted, using values of K and σ_n computed in a short simulation of the SDM using equations 3 and 4. It can be seen that there is a close correspondance between theoretical and simulated results. In figure 11 the experimental results are repeated for DS and TSC grouping types and almost identical noise figures are produced. These results indicate that the aliassing noise dominates over PWM intermodulation noise, and validates the conjecture that the introduced noise is independent of the grouping type.

In figure 12 the variation of the 2nd harmonic distortion with input level, for a SS PGM system with $L = 64, N = 8$ is compared to theoretical results. Results are presented for 1 kHz and 5 kHz input tones. The validity of the PGM model is again demonstrated by the close correspondance between experimental and theoretical results. The curves diverge where the harmonic tones reach the rising noise floor of the converter. Results are also plotted on the same graph for DS and TSC PGM, but here experimental comparisons are not possible because the tones fall below the noise floor produced by aliassing. In figure 13 third harmonic levels are plotted for the three modulation types.

Considerably lower levels of distortion are produced by DS and TSC modulation types, with TSC modulation offering the added advantage of having only odd order harmonics.

7.2 Closed Loop PGM

The feedback system described in figure 4 with $T_N(z) = (1 - z^{-1})^2$ and $T_E(z) = 1 - z^{-N}$ has been simulated. The baseband noise power for different N and L with the three grouping types is shown in figure 14 and it can be seen that grouping types offer similar performance. The gain in SNR obtained using feedback compared with open-loop for SS PGM is plotted in figure 15. As N increases the SNR advantage decreases. This is due to the delay in the feedback signal which reduces the effective suppression due to equation 21. Furthermore, the delay forces the input node error ($U(z)$ in figure 7) to be larger, causing the quantizer gain K to reduce and the noise power of the SDM to increase. These effects are plotted in figures 16. Consequently it is desirable to use low values of N and obtain high SNRs at low PRFs using low oversampling ratios with higher order loop filters. The restriction on N also suggests that TSC modulation is optimal as it achieves half the PRF of other modulation types with the same delay.

7.3 High Order PGM with Ideal and Non-ideal Power Switches

In this section the performance of two high order PGM systems are compared to a high order SDM system. Both ideal and non-ideal power switches are considered. The optimisation technique described in section 5 has been used in the design of three systems, detailed in table 1, with parameters $F_s = 44.1\ kHz$, $L = 128$.

Table 1 System parameters

System	Mod. Type	Grouping	NTF		ETF		Dither
			Order	$B_n(dB)$	Order	$B_e(dB)$	
A	SDM	-	4	104	-	-	rect PDF 0.4 p-p
B	PGM standard	TSC N=4	6	100	5	70	none
C	PGM collis detect.	TSC N=4	6	84	5	62.5	none

The filters have been designed to achieve stable operation with a 0.2 $p - p$ sinewave input. The differing NTF and ETF attenuations reflect the comparative stability of the three systems ($A > B > C$). System C is more difficult to stabilise because the additional quantizer inversions increase the instantaneous quantizer error, causing the quantizer gain to reduce.

Dither is used in the SDM system (A) to attenuate idle tones, however it has been found that the two PGM systems require no dithering, because the combination of the use of high order filters and the additional feedback loop tends to increase the complexity of the quantization error.

The systems have been simulated with a 1 kHz sinewave input of amplitude $0.2p - p$ using an ideal power switch and a switch with simulated rise and fall time mismatch of 1 nS. The baseband responses of the three systems are plotted in figures 17 and 18 for

the ideal and non-ideal cases, respectively. The SNRs are given in table 2. With ideal switches the SNR of the systems reflect the differing filter attenuations, with the SDM system achieving the best performance. With non-ideal switches the trend is reversed. and the PGM system with collision detection (C) offers the best SNR.

The range of PRFs exhibited by the systems have also been determined for a 1 kHz sinewave input varying in amplitude between zero and the modulator overload point. These results are given in table 2. The PRF of the SDM system is greater than the PGM systems and varies with input amplitude, causing noise intermodulation when non-ideal power switches are used (refer to section 6). The PRF of the standard PGM system is lower than the theoretical 705.6 kHz and also non-constant due to pulse group overloading. The PGM system with collision detection achieves a constant PRF, explaining its high tolerance to mismatched rise and fall times. In conclusion, the combination of a low and constant PRF makes the PGM system with collision detection the best candidate for power DACs.

Table 2 System Performance

System	SNR (dB) ideal	SNR (dB) non-ideal	PRF (kHz)
A	122.6	62.6	1697.4 - 2048.7
B	108.0	78.7	687.3 - 703.6
C	98.4	95.6	705.6

8 Conclusion

A digital signal processing technique has been described for power digital-to-analogue converters which offers lower bit clock rates and low distortion when compared to conventional pulse-width modulation converters. The system is based upon a modified SDM, where low PRFs are obtained by grouping together output pulses. The grouping has been modelled as a linear filtering, decimation and pulse-width modulation process. The noise and distortion introduced by this process are reduced by feedback. The performance potential has been increased using optimised high order filters. The considerably reduced bit-clock rates lead to the possibility of a full DSP solution without the need for discrete logic counters, making the system an attractive alternative to conventional PWM-based converters. Some of the effects of a non-ideal power switch have been reported and a modification to the PGM system has been briefly described, which considerably reduces the sensitivity to mismatched rise and fall times.

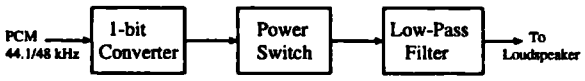


Figure 1: Block Diagram of Power Digital-to-Analogue Converter

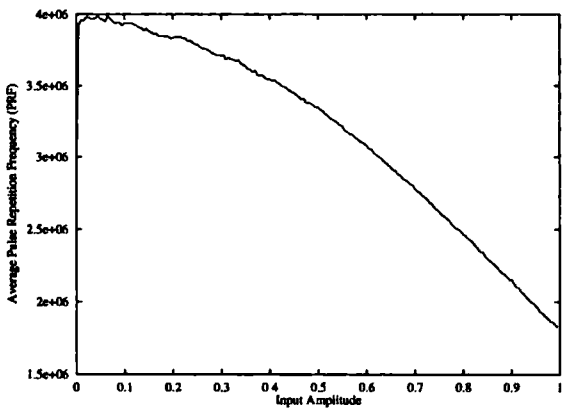


Figure 4: Average Pulse Repetition Frequency of 2nd order, L=256, Sigma-Delta Modulator

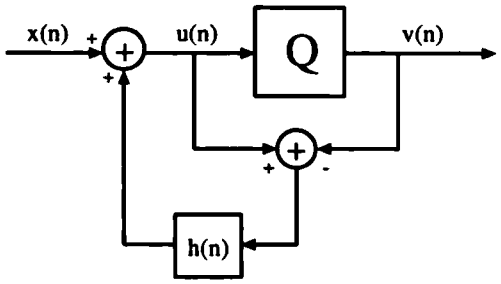


Figure 2: Noise-Shaper Topology of Sigma-Delta Modulator

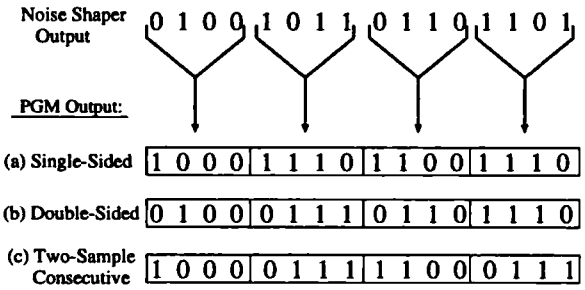


Figure 5: Pulse Grouping Schemes, $N = 4$

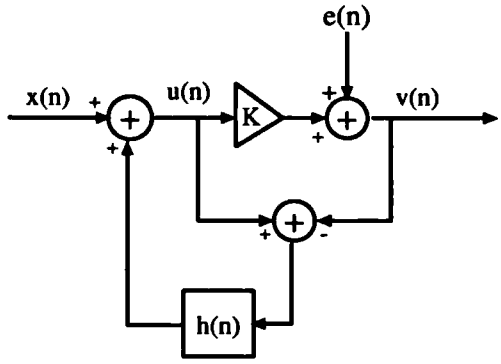


Figure 3: Quazi-Linear Model of Sigma-Delta Modulator

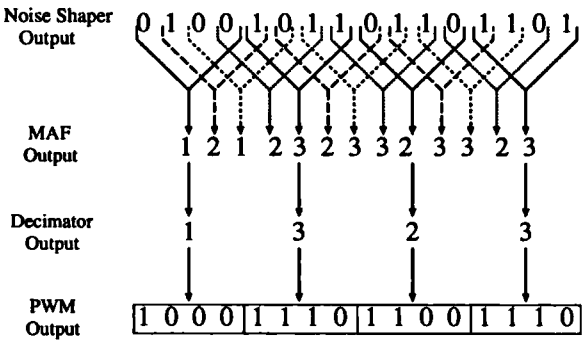


Figure 6: Operation of PGM Model for $N = 4$ (single sided grouping)

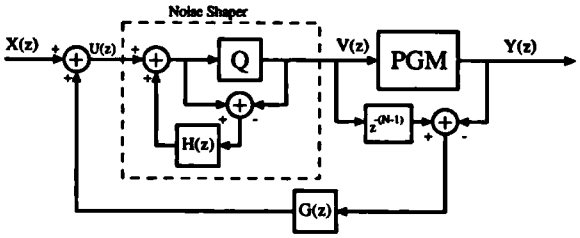


Figure 7: PGM With Feedback

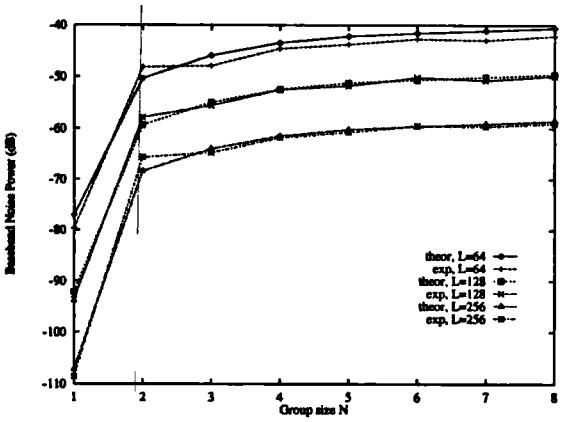


Figure 10: Noise Floor of Single-Sided PGM (experimental and theoretical)

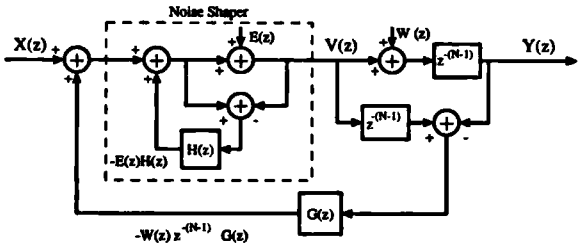


Figure 8: PGM Feedback Model

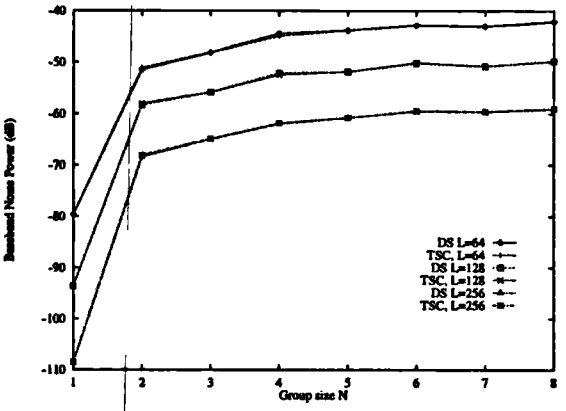


Figure 11: Noise Floor of Double-Sided and Two sample Consecutive PGM (experimental)

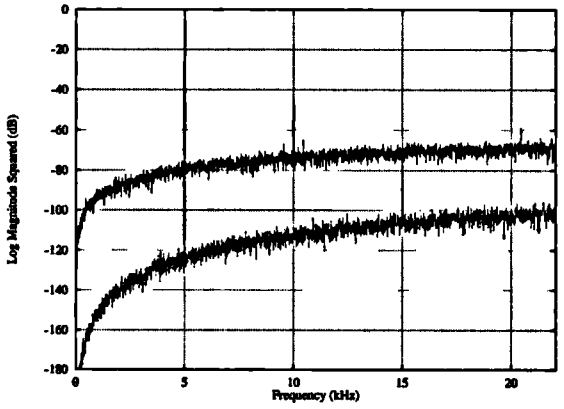


Figure 9: Baseband FFT of dithered single-sided PGM system (upper plot), $L = 64, N = 8$, SDM output (lower plot)

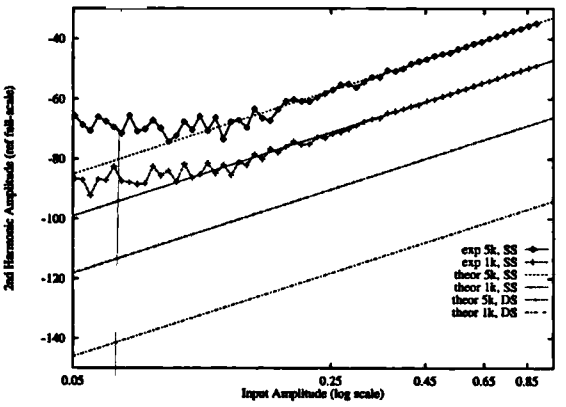


Figure 12: 2nd Harmonic Distortion for SS and DS PGM, 1kHz and 5kHz input (theoretical and experimental)

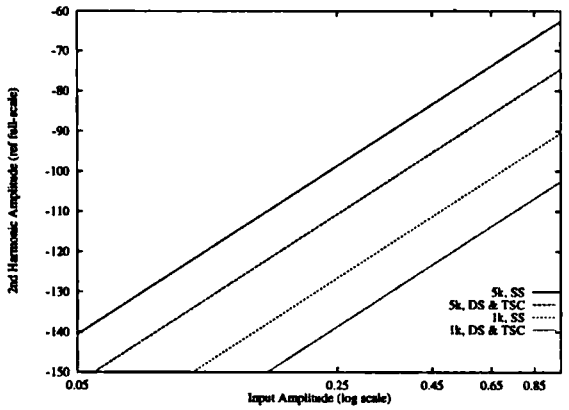


Figure 13: 3rd Harmonic Distortion for SS and DS sided and TSC PGM, 1kHz and 5kHz input (theoretical)

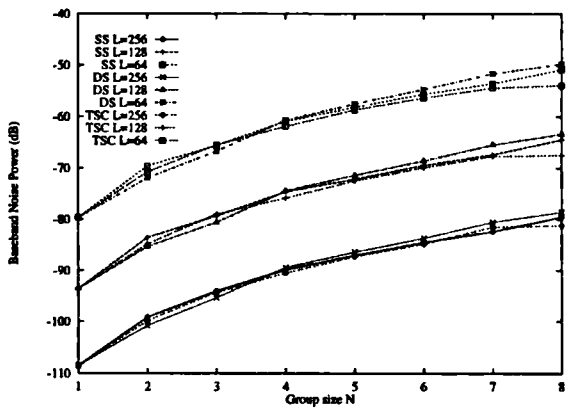


Figure 14: Baseband Noise Power of PGM with Feedback (experimental)

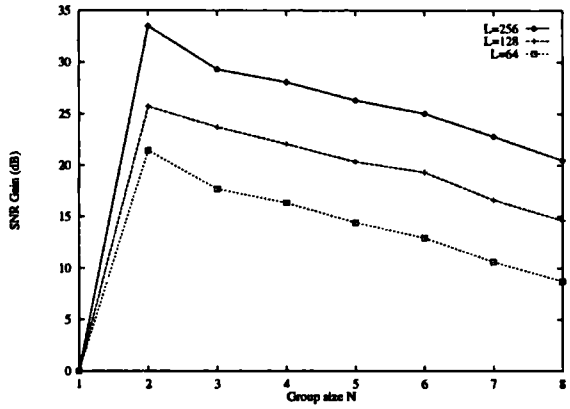


Figure 15: SNR Gain of SS PGM with Feedback (experimental)

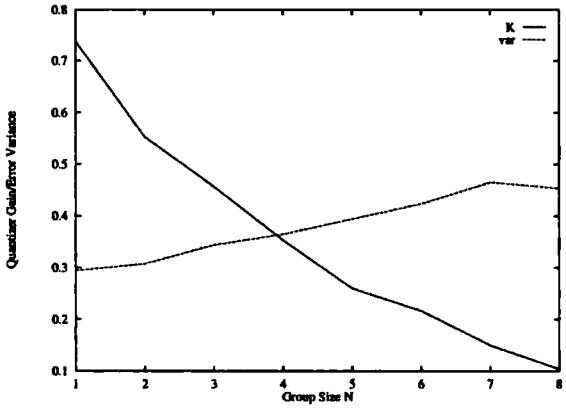


Figure 16: Quantizer Gain/Error Variance Variation with Group Size N (experimental)

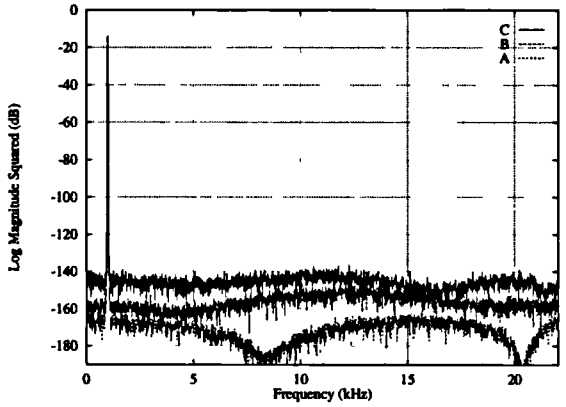


Figure 17: Baseband FFT of optimised high order SDM and PGM systems with ideal power switch

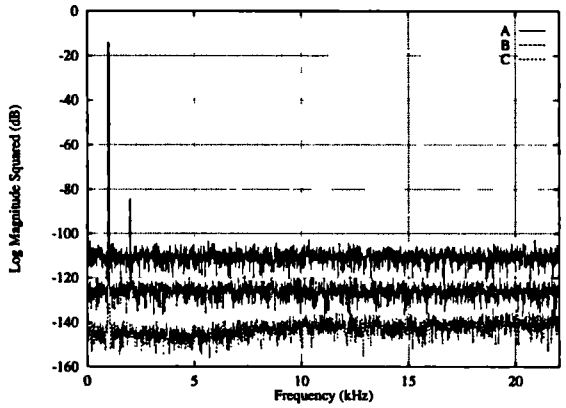


Figure 18: Wideband FFT of optimised high order SDM and PGM systems with non-ideal power switch

E.2 Power Switch Non-idealities

In this appendix we briefly consider how non-idealities in the power switch influences the performance of a $\Sigma\Delta$ power D-A converter. The aim is not to provide a detailed discussion of error mechanisms, but to establish problems which are peculiar to the bit-flipping system and may be avoided by appropriate algorithm design. Therefore, after a brief discussion of possible errors in the output stage, a particular error is concentrated on, which influences the linearity of $\Sigma\Delta$ D-A converters but does not influence the linearity of the competing PWM systems.

The linearity of the conversion to analogue for one-bit systems can be divided into two main types: amplitude errors and timing errors. Amplitude errors may occur, for instance, with a non-ideal power supply in which signal and load-dependent current drain causes the voltage supply to sag nonlinearly with signal amplitude. As 1-bit converters code the input signal by timing information, the design of the bit-flipping algorithm cannot directly influence this form of error, therefore amplitude errors will not be further discussed.

Timing errors can be subdivided into two main categories: Clock Jitter and Finite Rise and Fall Times.

E.2.1 Clock Jitter

Clock Jitter can be defined as the instantaneous timing deviation of the clock signal from its ideal sampling instants. Defining the jitter timing errors occurring on each sample as the jitter sequence $j(k)$, the sampling instants are [Ris94a]:

$$s(k) = kT_s + j(k) \tag{E.2}$$

where T_s is the sample period.

In one bit converters, the timing errors are observed after the sample and hold as deviations in the timing of each pulse transition from their ideal (uniformly sampled) times. In [Pau95], the influence of two forms of jitter is investigated for PWM D-A converters: non-accumulated and accumulated jitter. Non-accumulated jitter occurs where there is no correlation between adjacent jitter samples. This type of jitter may occur with a clock buffer stage which introduces timing uncertainty. Accumulated jitter occurs where timing errors accumulate between samples, and so a timing error on one sample displaces the timing of all subsequent samples.

Although the sensitivity of a 1-bit D-A converter is different to each of these forms of jitter [Ris94a], there are two similarities:

- Both forms of jitter are manifested as amplitude errors upon demodulation. and affect the noise performance of the system, and in the case of deterministic jitter, also the linearity.
- Errors are introduced only when a pulse transition occurs (since it is the transition times which are modified by the jitter).

The latter implies that the sensitivity of the converter to timing errors is related to the transition rate in the bitstream, and so reducing the transition rate will reduce the sensitivity to jitter. In [Ris94a] it is show that, in the case of non-accumulated jitter, the variance of the error sequence is approximately the product of the jitter variance and the variance of $y(k) - y(k-1)$, where $y(k)$ is the modulator output sequence. The variance of $y(k) - y(k-1)$ is inversely related to the average PRF of the bitstream, therefore the sensitivity to jitter is also inversely related to the average PRF.

For non-accumulated jitter, therefore, the sensitivity of PWM and $\Sigma\Delta$ modulator systems with the same average PRF will be approximately the same. Since jitter errors only occur on pulse transitions, regardless of jitter type, it is reasonable to assume that the sensitivity to accumulated jitter for PWM and $\Sigma\Delta$ systems with the same average PRF will also be approximately equal.

E.2.2 Unequal Rise and Fall Times

In this section the effects of finite rise and fall times are considered. In terms of linear errors, finite switching times cause the magnitude response to be attenuated at high frequencies [Cra93]. The main concern here is the effect of nonlinear errors i.e. how the linearity of the conversion to analogue is affected.

A simple model for the nonlinear artifacts of finite switching times has been proposed in [Ris94a], in which the change in area under the output waveform due to finite rise and fall times is expressed in the digital domain as an error impulse which occurs on every transition. This technique is based upon the method used in [Dun92] for modelling clock jitter errors with multibit signals. The main assumption made is that the errors occur at the sample instant, rather than being smeared with time - this assumption is more valid for smaller rise and fall times.

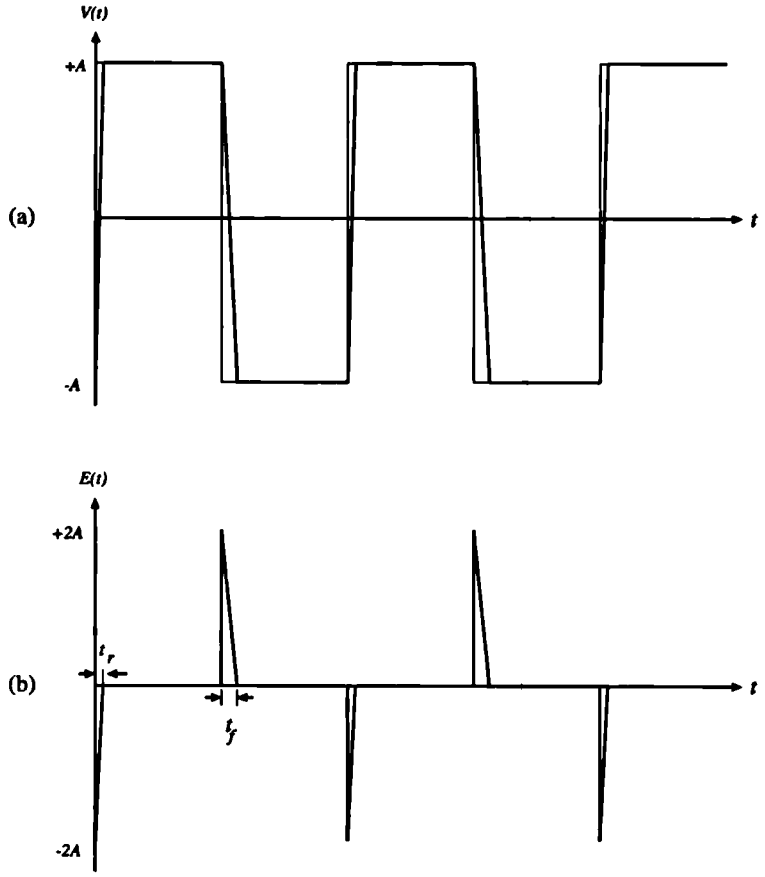


Figure E.2: Power switch output waveform (a) and (b) error waveform with finite rise and fall times.

In figure E.2, the error waveform for a square wave with rise and fall times t_r and t_f is shown. On every transition the error impulse is obtained by expressing the error area as a ratio of the oversampling period T_s [Dun92].

For an output signal of ± 1 the error impulse for a rising edge has strength

$$e_r = -2\left(\frac{t_r/2}{T_s}\right) = -\frac{t_r}{T_s} = -Lf_s t_r \tag{E.3}$$

and the error impulse for a falling edge is:

$$e_f = Lf_s t_f \tag{E.4}$$

A simplified approach to analysing the effects of these errors is to consider how the DC component varies with the input signal level [Ada89]. For a PWM system, the PRF is constant therefore the DC component is independent of input signal.

Unequal rise and fall times will result in a constant DC offset which is proportional to the PRF.

It has been shown in section 6.2.3 that the PRF of a $\Sigma\Delta$ modulator varies *nonlinearly* with input amplitude. For a DC input, a DC error will be introduced which varies nonlinearly with input amplitude. By modelling a bandlimited time-varying input as a slowly varying DC input, it can be seen that for a bandlimited input, a 'short-term' DC error will occur which varies nonlinearly with input level. In section 6.7.2 it will be shown by simulation that this variation gives rise to harmonic distortion [Ada89] and intermodulation (foldback) noise.

An error sequence is now derived for a 1-bit converter with rise and fall times t_r and t_f. For an ideal output sequence $y(k)$, a transition will occur whenever $y(k) \neq y(k-1)$. The following two expressions represent sequences which have the value of unity for a rising transition (i.e. $-1 \rightarrow +1$) and falling transition (i.e. $1 \rightarrow -1$) respectively.

$$\frac{1}{4}(y(k) - y(k-1))(1 + y(k)) \tag{E.5}$$

$$\frac{1}{4}(y(k) - y(k-1))(1 - y(k)) \tag{E.6}$$

These two expressions can be combined to derive the error sequence:

$$\epsilon(k) = \frac{1}{4}(y(k) - y(k-1))\{(1 + y(k))e_r + (1 - y(k))e_f\} \tag{E.7}$$

$$= \frac{Lf_s}{4}(y(k) - y(k-1))\{(1 + y(k))t_r + (1 - y(k))t_f\} \tag{E.8}$$

Defining the difference in rise and fall times as $\delta t = t_r - t_f$, equation E.8 can be re-expressed as:

$$\epsilon(k) = \epsilon_l(k) + \epsilon_n(k) \tag{E.9}$$

where

$$\epsilon_l(k) = \frac{Lf_s}{2}(y(k) - y(k-1)) \tag{E.10}$$

$$\epsilon_n(k) = \frac{Lf_s\delta t}{4}(y(k) - y(k-1))(y(k) - 1) \tag{E.11}$$

The term $\epsilon_l(k)$ represents a linear filtering operation. The term $\epsilon_n(k)$ represents the nonlinear component of the error and is equivalent to the error sequence derived

in [Ris94a]. The magnitude of this error depends on the difference between rise and fall times and the frequency of occurrence of the error is related to the bitstream transition rate. Due to the squaring of the $y(k)$ term, second harmonic distortion is likely [Ris94a].

Taking into account the nonlinear component only, the system output including this error source is:

$$y'(k) = y(k) + \epsilon_n(k) \tag{E.12}$$

This sequence has been modelled and simulated with the bit-flipping system and results are presented in section 6.7.2.

Although this model is very basic, it reveals that $\Sigma\Delta$-based D-A converters are sensitive to nonlinear errors arising from unequal rise and fall times. In contrast, the linearity of PWM D-A converters is entirely independent of the rise and fall time matching. To reduce the sensitivity of $\Sigma\Delta$ modulators to these errors, it is essential that the average PRF of the bitstream is independent of input amplitude. The algorithms presented in the preceding work attempt to achieve this by appropriate bit-flipping. The results in section 6.7.2 show that even when the measured average PRF is constant, nonlinear distortion and noise can still occur. The reason for this that the instantaneous PRF can vary with time, yet maintain an average value which is independent of input amplitude.

Example	Order	P_n	S_a
1	3	1	1
2	3	1	2
3	3	2	1
4	5	1	1
5	5	1	2
6	5	2	1
7	7	1	1
8	7	1	2
9	7	2	1

Table E.1: System parameters for alternation constraint model examples

E.3 Simulations of Alternation Constraint Modelling

In this appendix, simulation results are presented for the modelling of the alternation constraint as a delaying $\Sigma\Delta$ modulator (refer to section 6.4). The aim is to validate that the modulator with bit-flipping algorithm and alternation constraint S_a is identical to a modulator with S_a additional unit delays in the loop. A large number of simulations have been performed and the nine results presented here represent a cross section of these simulations. The simulation parameters are detailed in table E.1 and the wideband spectra are presented in figures E.3 to E.11. In each case the oversampling ratio is 64 and the input signal is a 7 kHz sinewave of amplitude $A_s = 0.15$. Each pair of simulations have identical spectra, indicating the validity of the model.

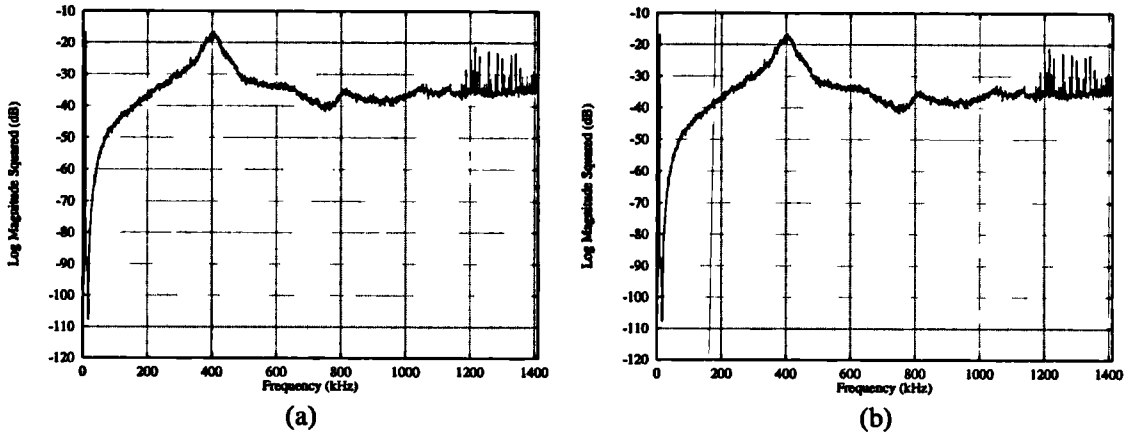


Figure E.3: System 1: (a) Alternation, (b) Delayed $\Sigma\Delta$

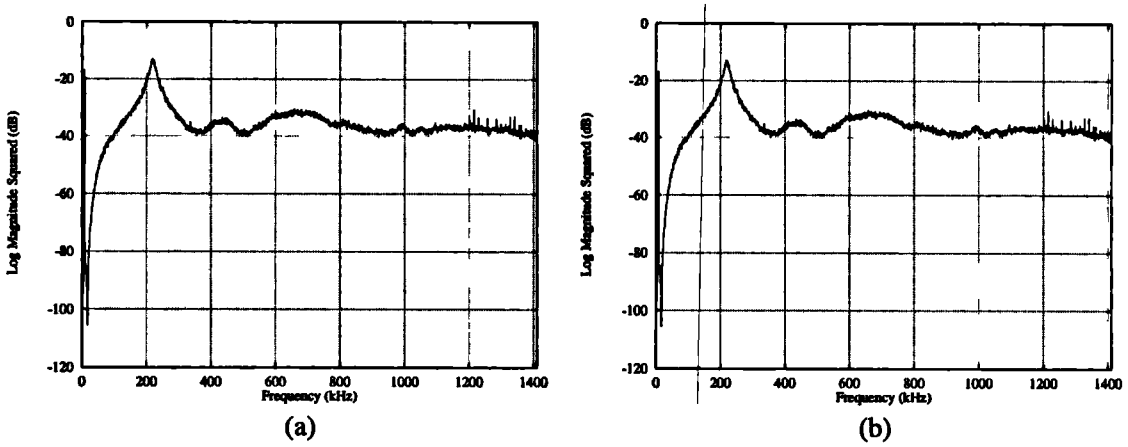


Figure E.4: System 2: (a) Alternation, (b) Delayed $\Sigma\Delta$

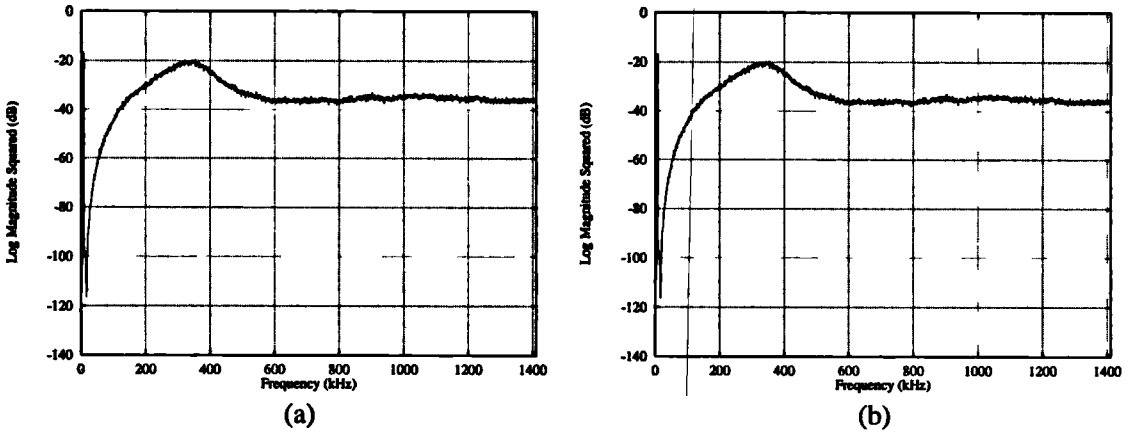


Figure E.5: System 3: (a) Alternation, (b) Delayed $\Sigma\Delta$

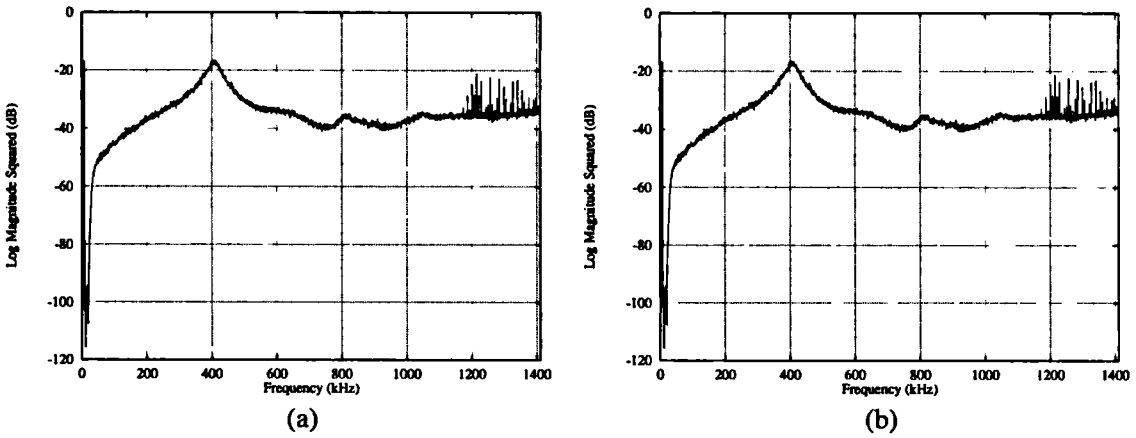


Figure E.6: System 4: (a) Alternation, (b) Delayed $\Sigma\Delta$

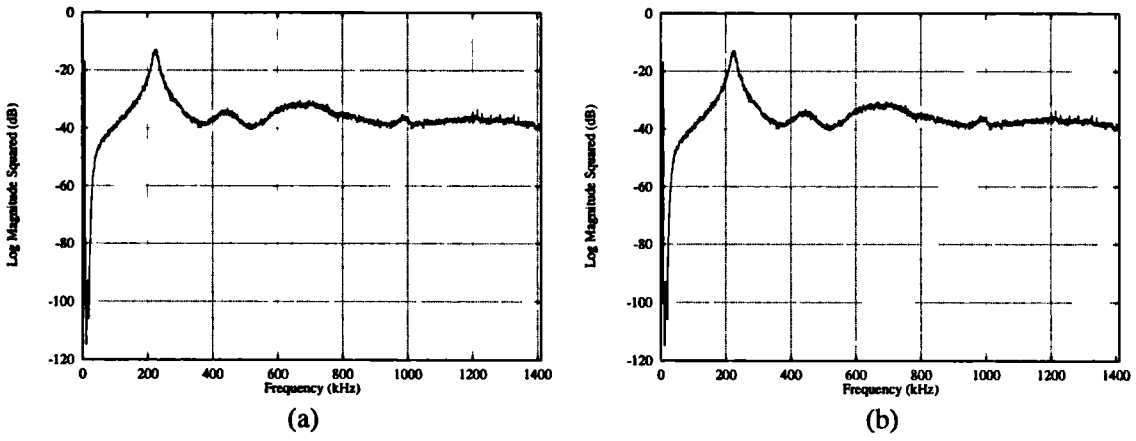


Figure E.7: System 5: (a) Alternation, (b) Delayed $\Sigma\Delta$

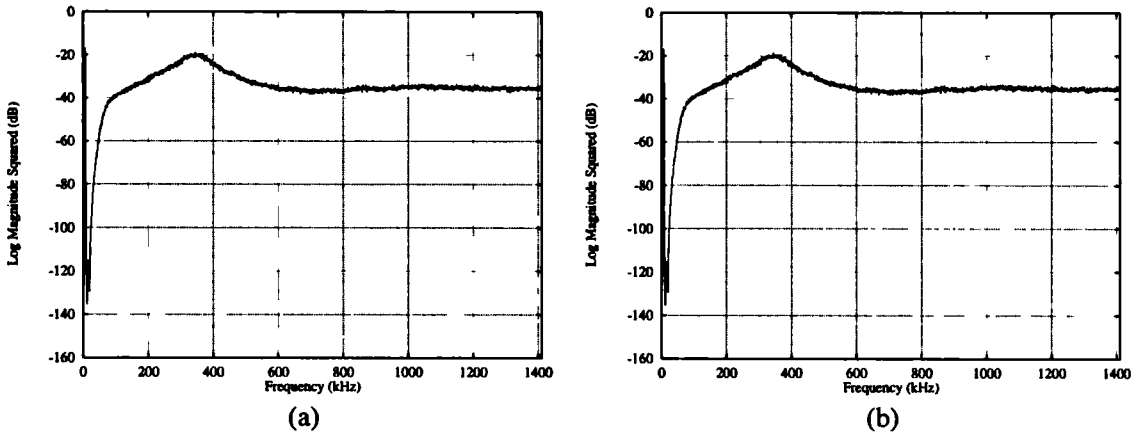


Figure E.8: System 6: (a) Alternation, (b) Delayed $\Sigma\Delta$

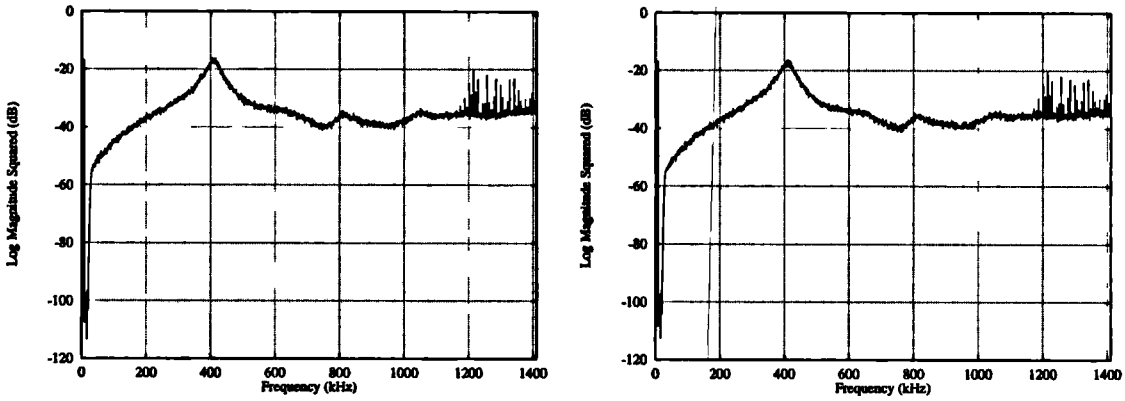


Figure E.9: System 7: (a) Alternation, (b) Delayed $\Sigma\Delta$

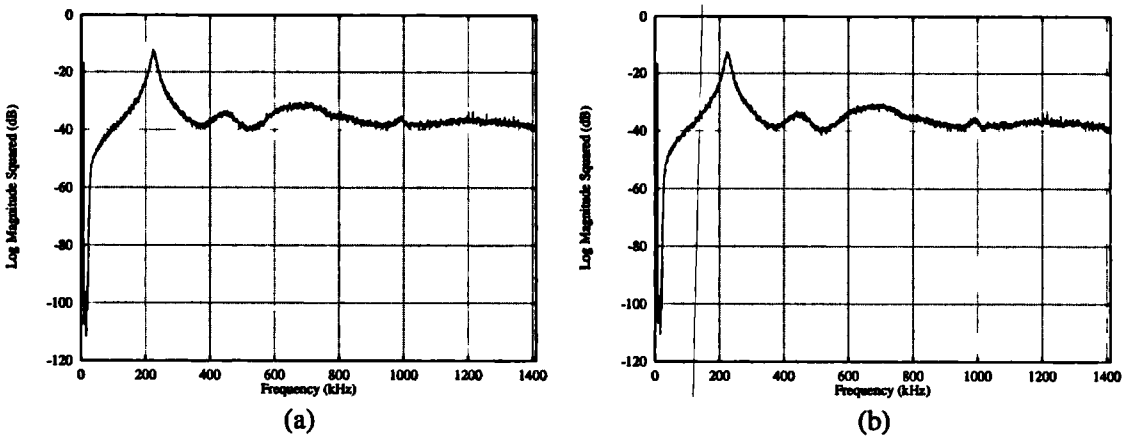


Figure E.10: System 8: (a) Alternation, (b) Delayed $\Sigma\Delta$

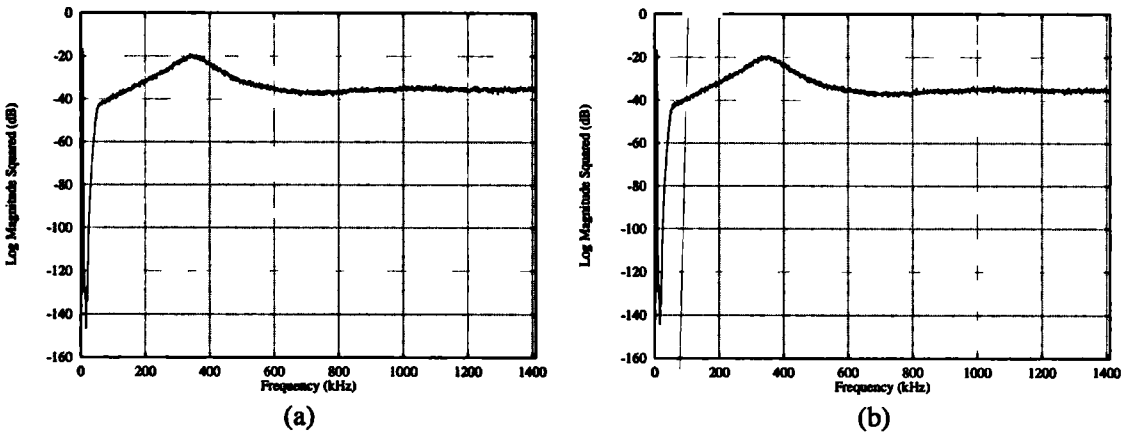


Figure E.11: System 9: (a) Alternation, (b) Delayed $\Sigma\Delta$

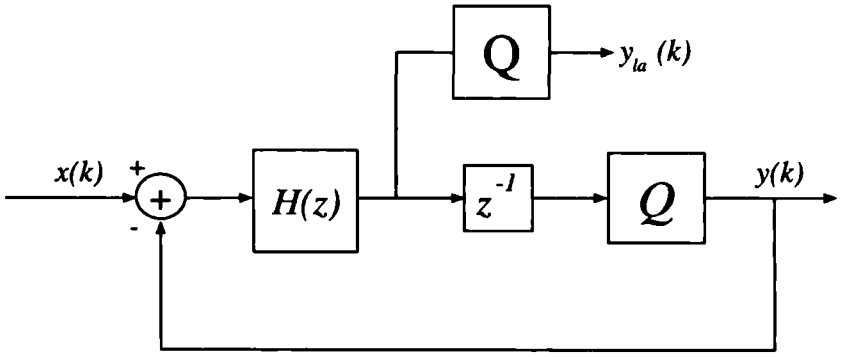


Figure E.12: Efficient Look-ahead Scheme with no Bit-flipping

E.4 Efficient Implementation of Look-Ahead Algorithm

It has been shown in section 6.5 that look-ahead may be used to improve the performance of the bit-flipping algorithm by preventing unnecessary algorithmic bit-flipping. In this appendix the look-ahead algorithm is examined and a method of efficient implementation is described. The principle is described in relation to the system with one-level of look-ahead, though the general principle also applies to the the system with two-levels of look-ahead.

The look-ahead algorithm increases the complexity of the system because the filter and quantizer output must be calculated again for every level of look-ahead whenever the quantizer state changes. It is important to note, however, that changing the state of the quantizer output only affects the filter states in a limited way and therefore it is possible to improve the efficiency of the look-ahead.

Starting from a standard modulator, the most efficient way of calculating the look-ahead is to remove the implicit delay from the loop filter and place it inside the loop. The quantized loop filter output is now an 'advanced' (look-ahead) version of $y(k)$ (figure E.12). This structure cannot be directly used if there is a bit-flipping operator (BFO) in the loop, since the look-ahead output will have already responded to the bit-flipping operation and due to a violation of causality it cannot be used to control the BFO.

The violation is solved by placing the BFO outside the loop. In this way, $y_{la}(k)$ always represents a prediction of $y(k)$ assuming no bit-flipping has occurred. If bit-flipping does occur, changes must be made to the loop variables so that their values

are correct *as if bit-occurred within the loop.*

The necessary changes to the loop-filter can be derived by considering the loop variables which have a dependency on $y(k)$. A direct form filter is used here (figure E.13), though the same principles can be applied to different filter structures.

$$\begin{aligned} v_o(k) &= x(k) - y(k) - b_1 v_1(k) - b_2 v_2(k) - ... \\ h_o(k) &= a_0 v_o(k) + a_1 v_1(k) + a_2 v_2(k) + ... \end{aligned} \tag{E.13}$$

If a decision is made to invert, the sign of $y(k)$ changes and the values of $h_o(k)$ and $v_o(k)$ are no longer correct:

With bit-flipping in sample k the correct values are:

$$\begin{aligned} v_{of}(k) &= x(k) + y(k) - b_1 v_1(k) - b_2 v_2(k) - ... \\ &= v_o(k) + 2y(k) \qquad \text{(E.14)} \\ h_{of}(k) &= a_0 v_{of}(k) + a_1 v_1(k) + a_2 v_2(k) + ... \\ &= h_o(k) + 2a_0 y(k) \qquad \text{(E.15)} \end{aligned}$$

Compensation is made by correcting these values as they propagate on the next sample instant using multiplexers in the filter structure and modulator loop.

$$v_1(k) = \begin{cases} v_o(k-1) & \text{no flipping} \\ v_o(k-1) + 2y(k-1) & \text{with flipping} \end{cases} \tag{E.16}$$

$$h_1(k) = \begin{cases} h_o(k-1) & \text{no flipping} \\ h_o(k-1) + 2a_0 y(k-1) & \text{with flipping} \end{cases} \tag{E.17}$$

These multiplexers are implemented in the loop filter and $\Sigma\Delta$ loop, as shown in figure E.13 and E.14.

The efficient one-level look-ahead algorithm only requires two extra low-resolution additions plus some control logic.

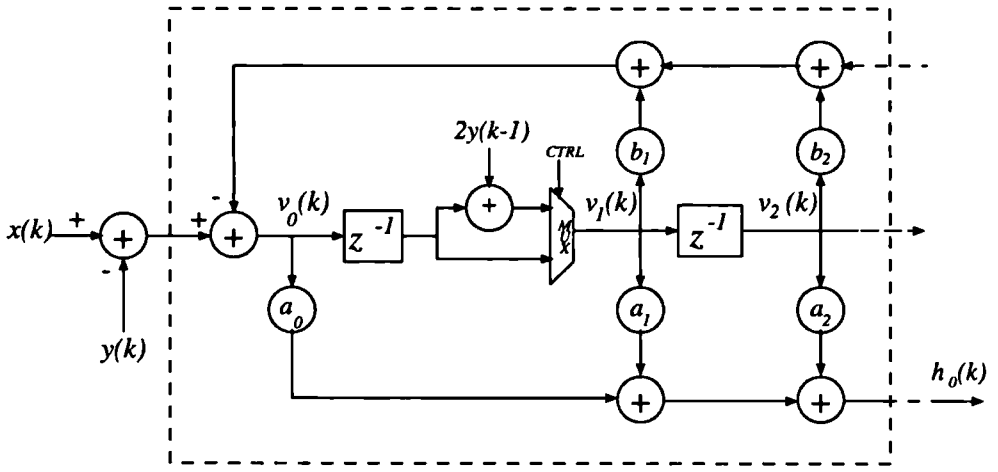


Figure E.13: Loop filter of efficient look-ahead scheme

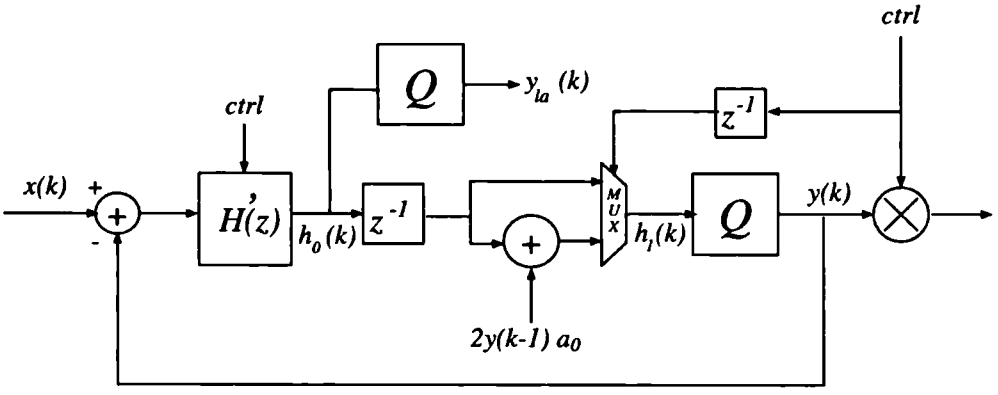


Figure E.14: Efficient look-ahead loop structure

References

[Ada84] R. W. Adams. Companded predictive delta modulation: A low-cost conversion technique for digital recording. *Journal of the Audio Engineering Society*, 32(9):63–77, September 1984.

[Ada89] R. W. Adams. An IC chip set for 20 bit A/D conversion. *Proceedings of the AES 7th International Conference: Audio in Digital Times*, pages 63–77, May 1989.

[Ada91] R. W. Adams, P. F. Ferguson, Jr., A. Ganesan, S. Vincelette, A. Volpe, and R. Libert. Theory and practical implementation of a fifth-order sigma-delta A/D converter. *Journal of the Audio Engineering Society*, 39:515–528, July-August 1991.

[Agr83] B.P. Agrawal and K.Shenoi. Design methodology for sigma-delta modulation. *IEEE Transactions on Communications*, COM-31(3):360 370, March 1983.

[Al-95] M. Al-Janabi, I. Kale, and R.C.S. Morling. Mash structures for bandpass sigma-delta modulators. In *IEE colloquium on Oversampling and sigma-delta Strategies for DSP, ref 1995/217*, November 1995.

[And93] M.A.E. Andersen. A new application for zero-current-switched full-wave resonant converters. *5th European Conference on Power Electronics and Applications*, 1993 September.

[Ard87] S.H. Ardalan and J.Paulos. An analysis of nonlinear behaviour in delta-sigma modulators. *IEEE Transactions on Circuits and Systems*, CAS-34(6):593 603, June 1987.

[Ard88] S.H. Ardalan. Analysis of delta-sigma modulators with bandlimited gaussian inputs. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1866–1869, 1988.

[Ath82] D.P. Atherton. *Nonlinear Control Engineering - Describing Function Analysis and Design.* Van Nostrand, 1982.

[Bai93] R.T. Baird and T.S. Fiez. Stability analysis of high order modulators for delta-sigma ADCs. In *IEEE International Symposium on Circuits and Systems*, pages 1361–1364, May 1993.

[Ben48] W.R. Bennet. Spectra of quantized signals. *Bell System Technical Journal*, 27:446–472, July 1948.

[Boo53] J.R.C Booton. The analysis of nonlinear control systems with random inputs. In *Proc. Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn*, April 1953.

[Can74] J. C. Candy. A use of limit cycle oscillations to obtain robust analog-to-digital converters. *IEEE Transactions on Communications*, COM-22(3):298–305, March 1974.

[Can76] J. C. Candy. Using triangularly weighted interpolation to get 13-bit PCM from a sigma-delta modulator. *IEEE Transactions on Communications*, COM-24:1268–1275, November 1976.

[Can85] J. C. Candy. A use of double integration in sigma-delta modulation. *IEEE Transactions on Communications*, COM-33(3):249–258, March 1985.

[Can86] J. C. Candy. Decimation for sigma-delta modulation. *IEEE Transactions on Communications*, COM-34(1):72–76, January 1986.

[Can92] J C. Candy and G. C. Temes, editors. *Oversampling Delta-Sigma Data Converters : Theory, Design and Simulation.* IEEE Reprints. IEEE Press, 1992. ISBN 0-87942-285-8.

[Cat88] F. Catthoor. SAMURAI: A general and efficient simulated annealing schedule with fully adaptive annealing parameters. *VLSI journal of Integration*, 6, 1988.

[Cha90] K.C.H. Chao, S Nadeem, W.L. Lee, and C.G. Sodini. A higher order topology for interpolative modulators for oversampling A/D converters. *IEEE Transactions on Circuits and Systems*, CAS-37(3):309–318, March 1990.

[Cho91] W. Chou. Sigma delta and multi-stage sigma-delta modulation with inside loop dithering. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1953–1956, May 1991.

[Chu84] S.Chu and C.S. Burrus. Multirate filter design using comb filters. *IEEE Transactions on Circuits and Systems*, 31, November 1984.

[Cra93] P. Craven. Towards the 24-bit DAC: novel noise-shaping topologies incorporating correction for the non-linearity in a PWM output stage. *Journal of the Audio Engineering Society*, 41(5):291–312, 1993.

[Cut60] C.C Cutler. U.S. Patent No. 2,927,962 (filed 1954), 1960.

[Dar90] T.F Darling and M.O.J. Hawksford. Oversampled analogue-to-digital conversion for digital audio systems. *Journal of the Audio Engineering Society*, 38(12):924–941, December 1990.

[Dav96] A.C. Davies. Periodic non-linear oscillations from bandpass sigma-delta modulators. *IEEE International Symposium on Circuits and Systems*, May 1996.

[Del92] D.F Delchamps. Quantization noise in sigma-delta modulators driven by deterministic inputs. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 425–428, March 1992.

[Dun92] C. Dunn and M.O. Hawksford. Is the AESEBU/SDPIF digital audio interface flawed ? *Presented at the 93rd Convention of the Audio Engineering Society, San Francisco*, Preprint 3360, October 1992.

[Dun94] C. Dunn and M.B. Sandler. A simulated comparison of dithered and chaotic sigma-delta modulators. *Presented at the 97th Convention of the Audio Engineering Society, San Francisco*, Preprint 3962, November 1994.

[Dun95] C. Dunn and M.B. Sandler. Efficient linearisation of sigma-delta modulator using single-bit dither. *Electronic Letters*, 31, 1995.

[Dun96] Dr Chris Dunn, April 1996. Private discussion.

[Eng94] A. Maynard Engebretson. Benefits of digital hearing aids. *IEEE Engineering in Medicine and Biology*, 13(2), April/May 1994.

[Fee95] O. Feely. Theory of lowpass and bandpass sigma-delta modulation. In *IEE colloquium on Oversampling and sigma-delta Strategies for DSP, ref 1995/217*, November 1995.

[Fie89] L.D. Fielder. Human auditory capabilities and their consequences on digital-audio converter design. In *Proceedings of the AES 7th International Conference: Audio in Digital Times*, pages 45–62, May 1989.

[Fri88] V. Friedman. The structure of limmit cycles in sigma delta modulation. *IEEE Transactions on Communications*, COM-36(8):972–979, August 1988.

[Gal93] I. Galton. One-bit dithering in delta-sigma modulator-based D/A conversion. In *IEEE International Symposium on Circuits and Systems*, pages 1310–1313, May 1993.

[Ger89] M.A. Gerzon and P.G. Craven. Optimal noise shaping and dither of digital signals. *Presented at the 87th Convention of the Audio Engineering Society, New York*, October 1989.

[Gol89] J.M. Goldberg and M.B. Sandler. The application of noise shaping for an all digital audio power amplifier. *Presented at the 87th Convention of the Audio Engineering Society, New York*, Preprint 2832, October 1989.

[Gol93] J.M Goldberg. *Signal Processing for High Resolution PWM-based Digital to Analogue Conversion.* PhD thesis, King's College, University of London, May 1993.

[Gol94] J.M. Goldberg and M.B. Sandler. New high accuracy pulse width modulation based digital-to-analogue converter/power amplifier. *IEE Proceedings on Circuits, Devices and Systems*, 141(4):315–324, August 1994.

[Goud89] A. Goudie. Idle tones in oversampling ADCs. *Presented at the 87th Convention of the Audio Engineering Society, New York*, Preprint 2882, October 1989.

[Gou94] F. Gourgue and M. Bellanger. A bandpass subsampled delta-sigma modulator for narrowband cellular mobile communications. In *IEEE International Symposium on Circuits and Systems*, pages 353–356, May 1994.

[Gra87] R.M. Gray. Oversampled sigma-delta modulation. *IEEE Transactions on Communications*, COM-35(5):481–489, May 1987.

[Gra89] R.M. Gray. Spectral analysis of quantization noise in a single-loop sigma-delta modulator with dc input. *IEEE Transactions on Communications*, COM-37(6):558–599, June 1989.

[Har78] F.J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 38(1), January 1978.

[Har92] S. Harris. How to achieve optimum performance from delta-sigma A/D and D/A converters. *Presented at the 93rd Convention of the Audio Engineering Society, San Fransisco*, October 1992.

[Haw90] M.O.J Hawksford. A comparison of two-stage 4th-order and single stage 2nd-order delta-sigma modulation in digital-to-analogue conversion systems. *IEE International Conference on Analogue to Digital and Digital to Analogue Conversion*, pages 148–151, 1990. Conference Publication No. 343.

[Haw92] M. O. J. Hawksford. Dynamic model-based linearization of quantized pulse-width modulation for applications in digital-to-analogue conversion and digital power amplification systems. *Journal of the Audio Engineering Society*, 40(4):235–251, 1992.

[He90] N. He, F. Kuhlmann, and A. Buzo. Double loop sigma-delta modulation with dc input. *IEEE Transactions on Communications*, COM-38(4):487–495, April 1990.

[Hei91] S. Hein and A. Zakhor. On the stability of interpolative sigma-dela modulators. *IEEE International Symposium on Circuits and Systems*, pages 1621–1624, June 1991.

[Hei93a] S. Hein. On the stablility of sigma-delta modulators. *IEEE Transactions on Signal Processing*, SP-41(7):2322–2348, July 1993.

[Hei93b] S. Hein. Tone supression in general double-loop sigma-delta modulators using chaos. *IEEE International Symposium on Circuits and Systems*, May 1993.

[Hen96] Dr R.K. Henderson, May 1996. Private discussion at International Symposium on Circuits and Systems, Atlanta.

[Hi-95] Hi-fi choice. *Felden Publications ISSN 0955 111*, 149:28–29, December 1995.

[Hi-96] Hi-fi news - comment. *Link House Magazines Ltd, ISSN 0142-6230*, 41(5):3, May 1996.

[Hio93a] R.E. Hiorns and M.Sandler. A modified noise shaper structure for digital PWM DACs. *Presented at the 95th Convention of the Audio Engineering Society*, Preprint 3767, October 1993.

[Hio93b] R.E. Hiorns and M.B. Sandler. Power digital-to-analogue conversion using pulse width modulation and digital signal processing. *IEE proceedings-G*, 140(5):329–337, October 1993.

[Hio94] R.E. Hiorns. *Digital Signal Processing and Circuit Design for PWM DACs.* PhD thesis, King's College, University of London, November 1994.

[Ino63] H. Inose and Y. Yasuda. A unity bit coding method by negative feedback. *Proceedings of the IEEE*, 51:1524–1535, November 1963.

[Iso91] J. Isoaho, H. Tenhunen, J. Heikkila, and L. Lipasti. High resolution DAC design based on FPGAs. In *Proc. Oxford International Workshop on FPGA Logic and Applications*, pages 343–352, September 1991.

[Ker96] S.M. Kershaw. *Sigma-Delta Bitstream Processors - Analysis and Design.* PhD thesis, King's College, University of London, August 1996.

[Klu92] J. Klugbauer-Heilmeier. A sigma delta modulated switching power amplifier. *Presented at the 92nd Convention of the Audio Engineering Society*, Preprint 3586, March 1992.

[Led3] R.C Ledzius and J. Irwin. The basis and architecture for the reduction of tones in a sigma-delta DAC. *IEEE Transactions on Circuits and Systems*, 40(7):429 439, July 1993.

[Lee87] W.L Lee and C.G. Sodini. A topology for higher order interpolative coders. *IEEE International Symposium on Circuits and Systems*, pages 459–462, May 1987.

[Lei90] S. P. Leigh, P. H. Mellor, and B. M. G. Cheetham. The implementation and performance enhancement of a completely digital power amplifier. *Proceedings of Institute of Acoustics*, 12(8):67 75, 1990.

[Lei91] S. P. Leigh P. H. Mellor and P B. M. G. Cheetham. Reduction of spectral distortion in class D amplifiers by an enhanced pulse width modulation sampling process. *IEE Proceedings-G*, 138(4):441–448, August 1991.

[Lip91] S.P. Lipshitz, J. Vanderkooy, and R. A. Wannamaker. Minimally audible noise shaping. *Journal of the Audio Engineering Society*, 39(11), November 1991.

[Mag95a] A.J. Magrath and M.B. Sandler. Efficient dithering of sigma-delta modulators with adaptive bit flipping. *Electronic Letters*, 31(11), May 1995.

[Mag95b] A.J. Magrath and M.B. Sandler. Efficient linearization of sigma-delta modulators with digital-domain dithering. *Presented at the 99th Convention of the Audio Engineering Society, New York*, Preprint 4105, October 1995.

[Mag95c] A.J. Magrath and M.B. Sandler. Non-linear deterministic dithering of sigma-delta modulators. In *IEE colloquium on Oversampling and sigma-delta Strategies for DSP, ref 1995/217*, November 1995.

[Mag95d] A.J. Magrath and M.B. Sandler. Power digital-to-analogue conversion using a sigma-delta modulator with controlled limit cycles. *Electronic Letters*, 31(4), February 1995.

[Mag95e] A.J. Magrath and M.B. Sandler. Power digital-to-analogue conversion using sigma-delta pulse inversion techniques. *Presented at the 99th Convention of the Audio Engineering Society, New York*, Preprint 4106, October 1995.

[Mag95f] A.J. Magrath and M.B. Sandler. Resolution enhancement and dither of sigma-delta modulator digital-to-analogue converters. *Electronic Letters*, 31(18), August 1995.

[Mag96a] A.J. Magrath and M.B. Sandler. Hybrid pulse width modulation / sigma-delta modulation power digital-to-analogue converter. *IEE Proceedings on Circuits, Devices and Systems*, 143(3):149–156, June 1996.

[Mag96b] A.J. Magrath and M.B. Sandler. Performance enhancement of sigma-delta modulator A-D and D-A converters using non-linear techniques. In *IEEE International Symposium on Circuits and Systems*, May 1996.

[Magu94] P.T. Maguire and Q. Huang. Quantizer gain in Nth-order sigma-delta modulator linear models: Its determination based on constant output power criterion. *IEEE International Symposium on Circuits and Systems*, pages 333–336, May 1994.

[Mok94] F. Mok, A.G. Constantinides, and P.Y.K Cheung. A flexible decimation filter for sigma-delta converters. *IEE Electronic Division Colloquium on Oversampling and Sigma-Delta Modulation, Digest No. 1994/083*, March 1994.

[Mat87] Y. Matsuya, K. Uchimura, and A. Iwata et al. A 16-bit oversampling A-to-D conversion technology using triple integration noise shaping. *IEEE Journal of Solid-State Circuits*, SC-22(6):921–929, December 1987.

[Mot93] M. Motamed, A. Zakhor, and S. Sanders. Tones, saturation, and SNR in double-loop $\Sigma - \Delta$ modulators. In *IEEE International Symposium on Circuits and Systems*, pages 1345–1348, 1993.

[Mot96] M. Motamed, A. Zakhor, S. Sanders, and T. Lee. Spectral characteristics of the double-loop $\Sigma\Delta$ modulator. In *IEEE International Symposium on Circuits and Systems*, 1996.

[Mut96] R.N Mutagi. Pseudo noise sequences for engineers. *IEE Electronics & Communications Engineering Journal*, 8(2):79–87, April 1996.

[Nag91] *Nag Fortran Library Mark 15.* The Numerical Algorithms Group Ltd., 1st edition, 1991. ISBN 1-85206-070-0.

[Nau87] P.J.A Naus, E.C. Dijkmans, E.F. Stikvoort, A.J. Mcknight, D.J. Holland, and W. Bradinal. A cmos stereo 16-bit D/A converter for digital audio. *IEEE Journal of Solid-State Circuits*, SC-22(3):390 394, June 1987.

[Nau88] P.J.A Naus and E.C Dijkmans. Low signal-level distortion in sigma-delta modulators. *Presented at the 84th Convention of the Audio Engineering Society, Paris*, Preprint 2584, March 1988.

[Nau91] P.J.A. Naus. Bitstream digital-to-analogue conversion for digital audio. MPhil, University College of Swansea, University of Wales, September 1991.

[Nor89] S.R. Norsworthy. A 14-bit 80 kHz sigma-delta A/D converter: Modelling, design and performance evaluation. In *IEEE Journal of Solid-State Circuits*, volume SC-24, pages 256–266, April 1989.

[Nor92] S.R. Norsworthy. Effective dithering of sigma-delta modulators. In *IEEE International Symposium on Circuits and Systems*, pages 1304 1307, 1992.

[Nor93] S.R. Norsworthy and D.A. Rich. Idle channel tones and dithering in delta-sigma modulators. *Presented at the 95th Convention of the Audio Engineering Society, New York*, Preprint 3711, 1993.

[Nor95] S.R. Norsworthy. Dynamic dithering of delta-sigma modulators. *Presented at the 99th Convention of the Audio Engineering Society, New York*, Preprint 4103, October 1995.

[Nut81] A.H. Nuttall. Some windows with very good sidelobe behaviour. *IEEE Transactions on Acoustics Speech and Signal Processing*, 29(1), February 1981.

[Opp89] A.V. Oppenheim and R.W. Schafer, editors. *Discrete-Time Signal Processing.* Prentice-Hall, London, 1989.

[Pau93] A. Paul and M.Sandler. Preliminary results of a 20-bit digital-to-analogue converter using pulse-width modulation. *Presented at the 95th Convention of the Audio Engineering Society*, Preprint 3766, October 1993.

[Pau95] A.C. Paul. *Linearisation of High Resolution Pulse Width Modulation based Digital-to-Analogue Converters.* PhD thesis, King's College, University of London, July 1995.

[Pla82] R.J. Van De Plassche and E.C. Dijkmans. A monolithic 16-bit D/A conversion system for digital audio. *Audio Engineering Society Premiere Conference, New York*, pages 54–60, June 1982. Collected Papers.

[Ped94] M.S. Pedersen. All digital power amplifier based on pulse width modulation. *Presented at the 96th Convention of the Audio Engineering Society, Amsterdam*, Preprint 3809, February 1994.

[Pro92] J.M Proakis and D.M. Manolakis. *Digital Signal Processing. Principles, Algorithms, Applications, Second Edition.* Macmillan Publishing Company, New York, 1992.

[Ree65] A.H. Reeves. The past, present and future of PCM. *IEEE Spectrum*, pages 58–63, 1965.

[Rib91] D.B. Ribner. A comparison of modulator networks for high order oversampled $\Sigma - \Delta$ analogue-to-digital converters. *IEEE Transactions on Circuits and Systems*, CAS-38(2):145–159, February 1991.

[Ris93] L. Risbo. Improved stability and performance from sigma-delta modulators using 1-bit vector quantization. *IEEE International Symposium on Circuits and Systems*, pages 1365–1368, May 1993.

[Ris94a] L. Risbo. FPGA based 32 times oversampling 8th order sigma-delta audio DAC. *Presented at the 96th Convention of the Audio Engineering Society, Amsterdam*, Preprint 3808, February 1994.

[Ris94b] L. Risbo. *Sigma-Delta Modulators - Stability Analysis and Optimisation.* PhD thesis, Technical University of Denmark, June 1994.

[Ris95] L. Risbo. On the design of tone-free $\Sigma - \Delta$ modulators. *IEEE Transactions on Circuits and Systems*, CAS-42:52–55, January 1995.

[Rit91] T. Ritoniemi, T. Karema, and H. Tenhunen. Modelling and performance estimation of sigma-delta modulators. In *IEEE International Symposium on Circuits and Systems*, pages 2705–2708, 1991.

[San83] M.B. Sandler. *Investigation by Simulation of a Digitally Addressed Audio Power Amplifier.* PhD thesis, University of Essex, October 1983.

[San95] Professor Mark Sandler, May 1995. Private discussion.

[Sch90] R. Schreier. An empirical study of high-order single-bit delta-sigma modulators. *IEEE Transactions on Circuits and Systems*, CAS-40(8):461–466, August 1990.

[Sch92] R. Schreier. Stability tests for single-bit sigma-delta modulators with second-order FIR noise transfer functions. *IEEE International Symposium on Circuits and Systems*, pages 1316–1319, 1992.

[Sch93] R. Schreier. Destabilizing limit cycles in delta-sigma modulators with chaos. In *IEEE International Symposium on Circuits and Systems*, pages 1369 1372, May 1993.

[Smi66] H.W. Smith. *Approximate Analysis of Randomly Excited Non-Linear Controls.* M.I.T Press Research Monograph No. 34, Cambridge, Mass :, 1966.

[Spa62] H.A. Spang III and P.M. Schultheiss. Reduction of quantizing noise by use of feedback. *IRE Transactions on Communication Systems*, pages 373–380, December 1962.

[Ste75] R. Steel, editor. *Delta Modulation Systems.* New York: Wiley, 1975.

[Sti88a] E. Stikvoort. Higher order one-bit coder for audio applications. *Presented at the 84th Convention of the Audio Engineering Society, Paris*, Preprint 2583, March 1988.

[Sti88b] E.F. Stikvoort. Some remarks on the stability and performance of the noise shaper or sigma-delta modulator. *IEEE Transactions on Communications*, COM-36(10):1157–1162, October 1988.

[Sto90] J.T. Stonick, S.H. Ardalan, and J.K Townsend. An improved analysis of $\Sigma\Delta$ modulators with bandlimited gaussian inputs. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1691–1694, April 1990.

[Tew78] S. K. Tewksbury and R. W. Hallock. Oversampled, linear predictive and noise-shaping coders of order N>1. *IEEE Transactions on Circuits and Systems*, CAS-25(7):436–447, July 1978.

[Tha62] G.J. Thaler and M.P Pastel. *Analysis and Design of Nonlinear Feedback Control Systems*. McGraw-Hill, New York:Electrical and Electronic Engineering Series, 1962.

[Thu94] A.M Thurston and M.J. Hawksford. Dynamic overload recovery mechanism for sigma-delta modulators. *Advanced A-D and D-A Conversion Techniques and their Applications, IEE Conference Publication No. 393*, pages 124–129, July 1994.

[Van84] J.V. Vanderkooy and S.P. Lipshitz. Resolution below the least significant bit in digital systems with dither. *Journal of the Audio Engineering Society*, 32(3):106–113, March 1984.

[Van87] J.V. Vanderkooy and S.P. Lipshitz. Dither in digital audio. *Journal of the Audio Engineering Society*, 35(12):966–973, December 1987.

[Van89] J.V. Vanderkooy and S.P. Lipshitz. Digital dither: Signal processing with resolution far below the least significant bit. *Proceedings of the AES 7th International Conference: Audio in Digital Times*, pages 87–96, May 1989.

[Vei96] B.R. Veillette and G.W. Roberts. FM signal generation using delta sigma oscillators. In *IEEE International Symposium on Circuits and Systems*, May 1996.

[Well89] D.R. Welland, B.P. Del Signore, E.J Swanson, T. Tanaka, K. Hamashita, S. Hara, and K. Takasuka. A stereo 16-bit delta sigma A/D converter for digital audio. *Journal of the Audio Engineering Society*, 37(6):476–483, June 1989.

[Yu92] J. Yu. *Design and Analysis of Fixed and Adaptive Sigma-Delta Modulators.* PhD thesis, King's College, University of London, September 1992.